Using test data to evaluate rankings of entities in large scholarly citation networks

by

Marcel Dunaiski

Dissertation approved for the degree of Doctor of Philosophy
in the Faculty of Science at Stellenbosch University

Computer Science Division,
Department Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Promoters:

Prof. W. Visser   Prof. J. Geldenhuys

April 2019

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

This dissertation includes five original papers published in peer-reviewed journals or books. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and, for each of the cases where this is not the case, a declaration is included in the dissertation indicating the nature and extent of the contributions of co-authors.

Date: April 2019

Declaration by the candidate:
With regard to the papers (published and unpublished) appended to this dissertation, the nature and scope of my contributions were as follows:

Nature of contribution	**Extent of contribution (%)**
All aspects of producing the papers	100

The following co-authors have contributed to the papers (published and unpublished) appended to this dissertation:

Name	**e-mail address**	**Nature of contribution**	**Extent of contribution (%)**
Willem Visser	wvisser@cs.sun.ac.za	Supervisory and editorial	50
Jaco Geldenhuys	jaco@cs.sun.ac.za	Supervisory and editorial	50

Signature of candidate: "Declaration with signature in possession of candidate and supervisors"

Date: 2018/12/03

Declaration by the co-authors:
The undersigned hereby confirm that

1. the declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-authors to the papers (published and unpublished) appended to this dissertation,
2. no other authors contributed to the papers (published and unpublished) appended to this dissertation, and
3. potential conflicts of interest have been revealed to all interested parties and that the necessary arrangements have been made to use the material in the papers (published and unpublished) appended to this dissertation.

Signature	**Institutional affiliation**	**Date**
"Declaration with signature in possession of candidate and supervisors"	Stellenbosch University	2018/12/03
"Declaration with signature in possession of candidate and supervisors"	Stellenbosch University	2018/12/03

Abstract

Using test data to evaluate rankings of entities in large scholarly citation networks

M.P. Dunaiski

Computer Science Division,
Department Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Dissertation: PhD (Computer Science)

March 2019

A core aspect in the field of bibliometrics is the formulation, refinement, and verification of metrics that rate entities in the science domain based on the information contained within the scientific literature corpus. Since these metrics play an increasingly important role in research evaluation, continued scrutiny of current methods is crucial. For example, metrics that are intended to rate the quality of papers should be assessed by correlating them with peer assessments. I approach the problem of assessing metrics with test data based on other objective ratings provided by domain experts which we use as proxies for peer-based quality assessments.

This dissertation is an attempt to fill some of the gaps in the literature concerning the evaluation of metrics through test data. Specifically, I investigate two main research questions: (1) what are the best practices when evaluating rankings of academic entities based on test data, and (2), what can we learn about ranking algorithms and impact metrics when they are evaluated using test data? Besides the use of test data to evaluate metrics, the second continual theme of this dissertation is the application and evaluation of indirect ranking algorithms as an alternative to metrics based on direct citations. Through five published journal articles, I present the results of this investigation.

Opsomming

Die evaluering van rangordes van entiteite in groot wetenskaplike sitasienetwerke deur die gebruik van toetsdata

M.P. Dunaiski

Rekenaarwetenskap Afdeling,
Departement van Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.

Proefskrif: PhD (Rekenaarwetenskap)

Maart 2019

Kern werksaamhede in die veld van bibliometrika is die formulasie, verfyning en verifikasie van maatstawwe wat rangordes vir wetenskaplike entiteite bepaal op grond van die inligting bevat in die literatuur korpus van die wetenskap. Aangesien hierdie maatstawwe 'n al belangriker rol speel in die evaluasie van navorsing, is dit krities dat hulle voortdurend en noukeurig ondersoek word. Byvoorbeeld, maatstawwe wat veronderstel is om die gehalte van artikels te beraam, moet gekorreleer word met eweknie-assesserings. Ek takel die evaluasie van maatstawwe met behulp van toetsdata gebaseer op 'n ander tipe objektiewe rangorde (verskaf deur kenners in 'n veld), en gebruik dít om in te staan vir eweknie-assesserings van gehalte.

Hierdie proefskrif poog om van die gapings te vul as dit kom by die evaluasie van maatstawwe met behulp van toetsdata. Meer spesifiek ondersoek ek twee vrae: (1) wat is die beste praktyke vir die evaluasie van rangordes vir akademiese entiteite gebaseer op toetsdata, en (2) wat kan ons leer oor die rangorde algoritmes en oor impak-maatstawwe wanneer ons hulle met die toetsdata evalueer? Buiten die gebruik van toetsdata, is daar 'n tweede deurlopende tema in hierdie proefskrif: die toepassing en evaluering van indirekte rangorde algoritmes as 'n alternatief tot maatstawwe wat direkte sitasies gebruik. Die resultate van my ondersoek word beskryf in vyf reeds-gepubliseerde joernaal artikels.

Preface

This is a dissertation by publication. It starts with a brief introduction to the topic of citation impact metrics, followed by five chapters which very briefly summarise five papers, and concludes with a synopsis chapter. The five papers themselves are appended, for the convenience of the reader, to the end of the dissertation. The bibliographies for each paper are self contained, and only the references cited in the introduction and the discussion sections of the dissertation are included in the dissertation's bibliography.

Contents

Chapter 1

Introduction

Citation metrics constitute an essential tool in scientometrics and play an increasingly important role in research evaluation (Bornmann, 2017). A large number of metrics are proposed every year for rating papers and authors which are often generalised and applied to aggregate levels for the evaluation of departments, institutions, or even countries. The paper by Mingers and Leydesdorff (2015) is an excellent overview of the field of scientometrics, review of citation impact metrics, and discussion about how metrics are used for research evaluation and policy decisions.

An important responsibility of the research community is to continuously scrutinise current metrics and validate whether they fulfil their intended purpose. According to Bornmann and Mutz (2015), situations should be created or found in empirical research in which a metric can fail to achieve its purpose. A metric should be regarded as only provisionally valid if these situations could not be found or realised. For example, metrics that are intended to rate the quality of papers should be assessed by correlating them with peer assessments (Bornmann and Marx, 2015). However, collecting direct peer-assessed test data is time-consuming and expensive. We therefore use a proxy for this assessment which comprises test data based on other ground truth provided by domain experts. More specifically, we collected five test data sets that comprise researchers or papers that have won awards or similar accolades of achievement that indicate their impact or influence in their respective academic fields. This forms the central theme of this dissertation, which focuses on paper and author ranking metrics and their evaluation with the use of appropriate test data. In this chapter, and with the future chapters in mind, we give the reader a brief overview of citation metrics and their evaluation. At the same time, we will try to put the topics that this dissertation touches upon into the context of this overview.

The terminology used in the literature varies occasionally but we tried to stay as consistent as possible. Impact metrics (also referred to as indicators) are methods that score papers, authors, or entities at more aggregated levels (i.e., journals, institutions, etc.). PageRank is an indirect impact metric that is also used to score such entities. However, we also refer to PageRank as an algorithm. We use the term ‘evaluation measure’ for methods that measure the ranking performance of metrics. For example, in Chapter 3, we evaluate evaluation measures in terms of how well they evaluate impact metrics.

1.1 Bibliographic databases

The two most widely used publication databases for bibliometric research are Web of Science (WoS) and Scopus. They can be accessed through subscription-based web interfaces which accommodate only small-scale and very basic citation analyses. Professional bibliometric centres, therefore, often have full access to the WoS and Scopus databases (Waltman, 2016). Google Scholar is another important publication database. However, large-scale citation analyses using Google Scholar are difficult since its data can only be accessed through its web interface. There exist differences in the coverage and quality between these databases but results of comparative studies are relatively transient and become outdated when the databases are updated. However, Google Scholar seems to have better coverage for conference proceedings and non-English journals compared to Scopus and WoS (Meho and Yang, 2007). Furthermore, Scopus has a broader coverage and higher citation counts across most fields compared to WoS (Waltman, 2016).

The type of analyses and experiments conducted for this dissertation are large-scale and require direct access to a database. Since we do not have access to the subscription-based databases, we use three different alternative databases that are freely available. The first is the Microsoft Academic Graph (MAG) database which is multi-disciplinary spanning most of academia. The second is the digital library of the Association for Computing Machinery (ACM) and contains papers from the field of computer science. The third is the Microsoft Academic Search database, which is the predecessor of MAG.

Recent studies indicate that the coverage and quality of MAG is comparable to other cross-disciplinary publication databases and usable for carefully designed citation analyses. Hug and Brändle (2017) find that MAG has comparable coverage of journal and conference papers and indexes more document types (e.g., dissertations) compared to Scopus and WoS. Compared to Scopus, they also find that MAG has more comprehensive coverage of book-related document types and papers from conference proceedings, but that it has slightly less journal articles. They argue that MAG is suited for evaluative bibliometrics in most fields, but that it shows similar sparseness in coverage as Scopus and WoS in the humanities, non-English publications, and open-access publications.

On the paper level, Hug and Brändle (2017) find that MAG has high correlations of citation counts with Scopus and WoS (Spearman coefficient of 0.73 for both). They also find that the publication year is correct for 89.5% of all publications and that the number of authors is correct for 95.1% of journal articles.

For the author level, Harzing and Alakangas (2017) use a sample of 145 authors in five broad fields (life sciences, sciences, engineering, social sciences, and humanities) and compare their average citation counts across Scopus, WoS, Google Scholar, and MAG. They find that MAG's average citation counts are higher than both Scopus and WoS for all fields. However, MAG still falls behind Google Scholar in the fields of engineering, social science, and humanities and they ascribe it to Google Scholar's better coverage of books and non-traditional research outputs. In Chapter 6, based on a non-random set of 513 researchers, we show that the citation counts of the authors between MAG and Google Scholar are highly correlated (Pearson coefficient of 0.91). A similar high correlation exists for the authors' publication counts (Pearson coefficient of 0.86). The coverage and quality of the ACM digital library has not been studied in depth. In Chapters 4 and 5 we show that citation counts of papers and authors based on the ACM database are a factor smaller compared to MAG, since it only includes citations from within the computer science domain. However, in Chapter 4 we show that citation counts of two non-random sets of test papers are relatively highly correlated (Pearson coefficient of 0.83 and 0.88). On the author level the correlation of paper counts (Pearson coefficient of 0.75) and citation counts (Pearson coefficient of 0.64) between the ACM and MAG database is much lower for the above mentioned 513 researchers.

1.2 A primer on citation metrics

Direct citation metrics Given a bibliographic database comprising some set of papers and paper cross-citations one may compute paper-level impact metrics. When additional publication information is available (e.g., author, journal, affiliation, or country information) these metrics may often be generalised for aggregated research units. On the paper level the simplest metric is to count a paper's number of citations. The problems and drawbacks of using citation counts as an impact metric have been discussed for many years (Garfield, 1979). The underlying problem is that not all citations count equally. This arises from varying citation cultures among the academic fields which result in field-specific citation densities (Lundberg, 2007; Radicchi *et al.*, 2008). The output rate of fields also varies over time (Bornmann and Mutz, 2015). Citation distributions are generally skewed and long-tailed, with many papers receiving only very few citations, while only a few papers receive many citations. Comparing papers in the long tail of citation distributions is difficult, since it is not obvious how many citations constitute a large enough number to signify a meaningful difference (Ioannidis *et al.*, 2016). Therefore, percentile-based metrics have been proposed where the citation score of a paper is rated in terms of its percentile in the citation distribution of the field or journal to which it belongs (Pudovkin and Garfield, 2009; Leydesdorff *et al.*, 2011; Bornmann *et al.*, 2013).

For aggregated research units more possibilities of defining metrics exist since a number of papers (i.e., citation distributions) for each unit may be compared. When we consider research group units, metrics may be classified into two categories: (1) metrics that are size-dependent and (2) metrics that are size-independent (Waltman, 2016). Size-dependent metrics measure the overall impact or productivity of research units. When additional papers are assigned to a unit, its score does not decrease. The total number of papers or the *total number of citations* of a unit's papers is the most basic indicator. We refer to total number of citations as 'citation counts', 'Citations' or 'CountRank'. As mentioned before, citation distributions tend to be highly skewed and may be unfairly influenced by a small number of highly cited papers. To address this problem, alternative indicators are based on the idea of counting the *number of highly cited papers* where a predefined citation count threshold determines whether a paper is counted as highly cited or not. The $i10$-index used by Google Scholar is one example and counts an author's papers that have received 10 or more citations. These types of metrics are usually defined to measure excellence.

Size-independent metrics usually measure the average performance per paper for research units. The *average number of citations per paper* simply computes the mean citation count over a research unit's complete set of papers. The journal impact factor (Garfield, 1972) is such a metric, which uses journals (publication venues) as research units and computes the mean number of citations received by the papers published in the journal (for a certain year). The analogous size-independent metric to 'number of highly cited papers' is the *proportion of highly cited papers* of a research unit. Again, a predefined citation count threshold determines whether a paper is considered as highly cited and thus is counted in the numerator when computing the proportion. Size-independent metrics are usually used to compare entities of different sizes (e.g., journals, research groups, institutions, etc.). However, it should be noted that even if these metrics have been normalised for differences among fields in citation density, these indicators are still sensitive to differences among fields in publication density (Waltman, 2016).

Data selection When computing metrics, choices about what data to include have to be made. In many cases, only papers from a certain time period are considered. In addition, based on the selected papers, only citations from a certain time period may be considered (citation window). When computing journal impact metrics, usually a citation window of only a few years is used. For example, the journal impact factor uses a citation window of two years. For the paper level, Wang (2013) states that the choice of a citation window entails a trade-off between the accuracy (larger citation windows) and the timeliness (smaller citation windows) of metrics. Abramo *et al.* (2011) claim, based on Italian science papers, that in all fields with the exception of mathematics, a citation window of two or three years is sufficient. However, according to Wang (2013) there is no generally applicable rule for choosing the size of citation windows. We analyse the impact that different citation window sizes have on paper-level metrics in Chapter 4 and show that they are important when evaluating their performances.

On the author level citation windows should also be considered. Abramo *et al.* (2012) find, based on all researchers at Italian universities in the sciences, that a citation window of one year is acceptable for ranking authors and that for physics, biology, and medicine the ranking variation compared to larger citation windows is negligible. Similarly, Costas *et al.* (2011) show that delayed recognition of papers does not influence the rankings of individual researchers and therefore they are not penalised when shorter citation windows are used. In Chapter 5, we evaluate the field bias and ranking performance of impact metrics on the author level and show that the relative performances of metrics depend on the citation window size.

After citation windows are defined, additional papers or citations may be excluded. For papers the data selection criteria are mostly based on the document type (journal article, proceedings paper, editorial, etc.), language, and journal type. The decision mostly depends on the research question or is limited by the available data. For document types, it is sometimes necessary to exclude certain editorials for fair comparisons between researchers when using size-independent metrics since editorials tend to be cited less frequently than regular articles or papers from conference proceedings. In terms of language, non-English language papers should be excluded when computing institutional-level size-independent metrics, since they are cited less on average and would negatively bias institutions in which researchers publish frequently in their own language (van Raan *et al.*, 2011). For the MAG database, differentiation between document types and language is not possible since that information is not provided.

Self-citations may be defined at different aggregation levels and also excluded. For example, journal self-citations (i.e., a paper in a journal citing another paper in the same journal) are excluded from the Eigenfactor metric computation (Bergstrom *et al.*, 2008), while the journal impact factor includes them. However, author self-citations are most commonly investigated. Author self-citations are usually defined as citations for which the citing and the cited paper have at least one author in common (Aksnes, 2003). However, author self-citations can be classified into two types. The first type are the above-mentioned self-citations but renamed *co-author self-citations* (Costas *et al.*, 2010). Removing these type of self-citations, removes a potential citation for each co-author of the cited paper even if some of these co-authors are not on the citing paper. The second type are *author self-citations* and removing them only removes a citation for the authors that appear on both the citing and cited papers. On the paper-level, only one type of author self-citation exists which is the *co-author self-citation*. This is because a citation to a paper is either included or excluded based on the selected citation inclusion criterion. An alternative is to assign fractional citations to papers but this is only possible if author information is available. In Chapter 2 we investigate the impact in ranking performance of excluding author self-citations when counting citations on the author level. In Chapter 6 we distinguish between author-self citations and co-author self-citations and show that the distinction is important.

Counting methods Scientific production is becoming increasingly collaborative (Larivière *et al.*, 2014). The number of multi-authored papers and the average number of co-authors on papers increases in all fields (Wuchty *et al.*, 2007). The rates of collaboration on country and institution levels and the percentage of multi-authored papers however still differ among fields (Gazni *et al.*, 2011). In high energy physics and some biomedical fields hyperauthorship (papers with more than 100 authors) is nowadays common which also has implications for author accountability and how to apportion credit (Cronin, 2001).

Citation metrics usually allocate the full credit of a paper to each co-author. This method is named 'full counting' and has the effect that a citation to a multi-authored paper is counted multiple times. Instead of using full counting, one may divide a paper's credit equally among its co-authors. This is called 'fractional counting'. The main problem with fractional counting is that it might also not be fair.

The question of how to distribute the credit of papers among co-authors more appropriately has been discussed extensively and many methods of counting multi-authored papers have been proposed. Ideally, one would like to use perfect information about each author's contribution to a paper, but Ajiferuke *et al.* (1988) show that even interviewing the authors directly may be unreliable. Therefore, the information used as provided on the papers or to the publishers must be used. Trueba and Guerrero (2004) state three principles that should be followed when distributing scores between co-authors. The value of a paper should be shared between authors, divided among authors, and the first author should be credited more than the later authors on the paper. It should be noted that in some disciplines the conventions of author ordering are different. For example, if the ordering of author names is alphabetical, then assigning scores based on author positions would be inappropriate. For example, alphabetical ordering is common in mathematics, economics and high energy physics (Waltman, 2012).

Additional counting methods have been proposed that are based on the position of author names on papers. For example, one may only count the first-named authors (straight counting) or the corresponding authors of papers. Alternatively, one may share the credit with decreasing portions based on author positions (Howard *et al.*, 1987; Assimakis and Adam, 2010). At the level of individual authors, Lindsey (1980) shows that fractional counting is preferable and less biased towards authors of multi-authored papers compared to full counting. Lange (2001) shows that the correlation between straight counting and full counting is higher for established (senior) researchers than for junior researchers with fewer published papers. In Chapter 6 we evaluate various counting methods for multiple impact metrics based on test data comprising well-established researchers. We found that fractional counting is the best method for ranking well-established authors independent of which impact metric is used.

1.3 Indirect citation metrics

Indirect metrics not only take the direct citations into account to compute scores but also the indirect impact of citations through reference chains. Most indirect metrics are recursively defined and consider the entire network structure of citation graphs. The idea of recursively defining impact metrics originates with Pinski and Narin (1976). They applied it to academic citation networks to compute importance scores for journals to address the limitation that all citations are valued the same. The rationale of applying indirect metrics to citation networks is that citations from influential papers should count more than citations from unimportant papers. The advent of the Internet and the introduction of *PageRank* (Brin and Page, 1998) has led to renewed interest in recursively defined citation metrics.

Intuitively, PageRank simulates a process in which random walkers are placed on papers and follow citations to other papers. This continues until they are teleported to new random papers controlled by a teleportation probability $(1 - \alpha)$, where α is a parameter of PageRank called the 'damping factor'. If the random walkers reach papers without outgoing references, they restart their searches with new random papers.

Two aspects of this process are important to consider when applying PageRank on paper citation networks. Firstly, the damping factor has to be chosen carefully since it controls the average path lengths of the random walkers (Chen *et al.*, 2007; Ding *et al.*, 2009). And secondly, a personalisation vector controls the initial placement of the random walkers and their restarts (Nykl *et al.*, 2015; Fiala and Tutoky, 2017). It can be initialised, for example, to skew the probabilities towards papers in a certain field or papers published during a certain time period.

PageRank on the paper citation networks is inherently biased towards older papers due to the time-directed nature of the graphs (Dunaiski, 2014, p.97). In other words, the random walkers tend to quickly move towards older nodes in the network and disregard younger papers disproportionately (Chen *et al.*, 2007). Variations of PageRank have been proposed to overcome this drawback. Walker *et al.* (2007) proposed a PageRank-like iterative algorithm which simulates the dissemination of the random walker across the citation network, while taking into account that researchers typically start their search at recently published papers. They define a time decay parameter τ which exponentially biases the random walkers towards younger papers when they start or restart their searches. In addition, at each iteration, the probability that the random walkers are satisfied with their search increases quadratically.

Hwang *et al.* (2010) also use a time decay parameter τ to exponentially penalise older papers through the personalisation vector in PageRank. The final personalisation vector comprises the product of these values and the journal impact factor at which the papers are published. Similarly, Dunaiski and Visser (2012) use a time decay parameter for the personalisation of PageRank. In addition, each edge in the network is weighted inversely by the time difference between the citing and the cited paper. In other words, references to older papers are exponentially penalised. The last variant we want to mention here is SceasRank (Sidiropoulos and Manolopoulos, 2005). SceasRank is defined in such a way that direct citations contribute most to papers' scores. Furthermore, citations from recently published papers add more value to the score of recently published papers.

Since PageRank may be computed on any network, paper citation graphs may be collapsed to aggregated levels such as author citation graphs (Ding *et al.*, 2009; West *et al.*, 2013), co-authorship (or collaboration) networks (Liu *et al.*, 2005; Fiala *et al.*, 2008), or journal cross-citation graphs (Bollen *et al.*, 2006). We refer the reader to the paper by Fragkiadaki and Evangelidis (2014) in which citation graphs at different levels are defined and many recursively defined citation impact metrics are reviewed. Nowadays, citation scores based on PageRank-like algorithms are used by the WoS under the name Article Influence score (Bergstrom *et al.*, 2008), while Scopus reports the SCImago journal rank (SJR) metric (González-Pereira *et al.*, 2010).

An alternative indirect metric, proposed by Giuffrida *et al.* (2018), also takes the impact of citing papers into consideration. However, differently to PageRank, it only considers two citation levels and transfers less of the citing paper's impact to the cited paper, by implementing the restriction that the gain through citations from papers with high scores should not be more than 1 (i.e., two citations cannot count less than one).

In Chapter 2 we evaluate various PageRank variants in terms of ranking performance on the paper level while taking the damping factor into consideration. We also analyse PageRank's ranking bias in terms of fields and time while taking the damping factor and different personalisation strategies into account (Chapter 4). On the author level we evaluate PageRank with different personalisation strategies, counting methods, and damping factors (Chapter 6) and analyse its field bias (Chapter 5).

1.4 Metric normalisation

One of the key principles of bibliometrics is that entities from different fields should not be directly compared to each other through metrics that are based on pure citation counts. This stems from the observation that citation densities (average number of citations per paper) vary between academic disciplines (Lundberg, 2007) with field-dependent changes over time (Bornmann and Mutz, 2015). When comparing papers from different years, one should also account for the amount of time papers have had to accrue citations. For example, a paper published in 2000 with 40 citations should not necessarily be considered more impactful than a paper from 2010 with 10 citations, if the average citation count per paper in 2000 was 80 while in 2010 it was 10. Therefore, citations should not be treated equally.

The goal of incorporating normalisation into citation metrics is to try and account for the biases introduced through these varying citation potentials. In most cases, one corrects for field differences, publication years, and sometimes also the document type (e.g., article, review, letter) of papers (Waltman, 2016).

Publication years are usually associated with papers in publication databases and are unambiguous. To categorise papers into fields or disciplines is however not as trivial. In the past, fields have been categorised on the basis of journals or library categories. The problem is that fields and especially broad disciplines are not isolated. Generally, within-field citations are denser than between-field citations, however, between-field citations are becoming more common nowadays (Silva *et al.*, 2013). Furthermore, when only differentiating papers at the most aggregate level (i.e., disciplines), heterogeneities in the subfields' citation densities might be disregarded (van Leeuwen and Calero Medina, 2012). This is more problematic for recursively defined metrics compared to citation counts (Waltman *et al.*, 2011*c*). An agreement about the optimal classification scheme has yet to be reached (Zitt *et al.*, 2005; Adams *et al.*, 2008; Colliander and Ahlgren, 2011).

Alternatives to defining fields a priori exist. Fields may be defined dynamically based on the citation network structure (Rosvall and Bergstrom, 2011) or the semantic relatedness of papers' contents. Hutchins *et al.* (2016), for example, use a paper's co-citation network to define its field. The rationale is that if papers that are cited together by another paper, they belong to the same topic since they were relevant in producing the new paper. However, this does not hold for all citations since some work, such as statistical methods, is inherently of such a nature that it attracts citations from multiple, generally unrelated fields (Silva *et al.*, 2013).

Normalising over ill-defined fields may also lead to undesirable situations. For example, when a paper's field is defined by its co-citations and it receives new citations from a remote field it may indicate an increase in importance. However, as Waltman (2015) points out, if the remote field has a high citation density, normalisation may lead to a decrease rather than an increase in the paper's score with the newly acquired citations. Similarly, defining categories by individual journals (Pudovkin and Garfield, 2009) is also problematic and normalisation over such categories has to be well-justified. For example, the top paper from a prestigious journal should probably not be scored the same as the top paper from an obscure (generally under-cited) journal.

For the papers accompanying this dissertation in which we consider fields, we use two different categorisation schemes. On the ACM database, we use the ACM classification system (ACM, Inc., 2017) which consists of library-like categories that are fine-grained and author-chosen. For the MAG database, we use its field categorisation scheme which is based on the semantic information contained in papers' keywords, titles, and abstracts (Microsoft, 2017). However, the categorisation scheme of MAG for fine-grained topics is too noisy for citation analyses (Hug *et al.*, 2017). For our experiments we therefore only use the top-level fields (i.e., disciplines) in the MAG database (Vaccario *et al.*, 2017).

On the paper level, the basic idea of metric normalisation is to compute the expected citation scores for fields and use these to rescale the corresponding paper scores. This is also referred to as cited-side normalisation since the cited papers are the focus of the normalisation step. Typically, the normalised citation score of a paper p is the ratio of its actual score $S(p)$ and the expected score of its field μ (i.e., $S(p)/\mu$). Mariani *et al.* (2016) show that using the ratio for rescaling strongly depends on the age of papers and use the z-score (i.e., $(S(p) - \mu)/\sigma$) for normalisation where σ is the standard deviation of the paper scores of the associated field. This rescaling approach was initially proposed by Lundberg (2007) but where citation counts are first transformed with the natural logarithm (i.e., $(\ln(S(p) + 1) - \mu_{ln})/\sigma_{ln}$).

When also accounting for varying citation densities over years, time normalisation based on calendar years produces noisy results for the most recently published papers (Ioannidis *et al.*, 2016). Parolo *et al.* (2015) show that for all ages, the number of papers is a better indicator to capture the role of time in academic citation networks than actual time. Therefore, citation scores of papers may be rescaled by only considering the closest papers that were published around the same time, independent of the actual calendar years (Mariani *et al.*, 2016). We focus on paper-level normalisations in Chapter 4.

For aggregated research units, normalised metrics may be defined based on the expected paper scores or directly on the rescaled paper scores. For example, given a set of papers for an author, the two possible approaches for computing the *normalised average citation score* are:

1. the ratio of the sum of the papers' actual scores and the sum of the papers' expected scores

2. the average of the normalised paper scores (i.e., the papers' individual ratios)

The former is called the *ratio of sums/averages* (or 'globalised') approach, while the latter is called the *average of ratios* (or 'averaged') approach (Egghe and Rousseau, 1996; Waltman, 2016). Opinions about which approach is more appropriate for bibliometrics vary. Some argue that the averaged approach is better suited (Lundberg, 2007; Waltman *et al.*, 2011*a*), while others argue for the globalised approach (Moed, 2010; Vinkler, 2012).

Waltman *et al.* (2011*b*) show that the difference is very small at the country level and for large institutions, but that the difference is somewhat larger for research group and journal levels. Larivière and Gingras (2011) find that the difference at any aggregation level is statistically significant and that it depends on the unit's number of papers. They show that the difference between these two approaches is greater for departments than for individual researchers. They argue that this is because the diversity of topics in which departments publish is generally greater than that of individuals. In Chapter 5 we analyse the difference between the averaged and globalised aggregation approaches on the author level. We use various paper-level normalised impact metrics and evaluate the different aggregation variants in terms of field bias and ranking performance on the author level.

An alternative to cited-side normalisation is citing-side normalisation (also referred to as source-normalisation) as initially proposed by Zitt and Small (2008). The argument for citing-side normalisation is that a large factor of varying citation densities between fields is caused by different average references list lengths in fields. Opinions differ concerning which normalisation approach is better for accounting for field differences (Waltman and van Eck, 2013; Radicchi and Castellano, 2012). In this dissertation we do not directly evaluate citing-side-normalised citation metrics. However, the PageRank algorithm incorporates citing-side normalisation to some degree: it normalises the impact of each individual citation by the number of outgoing references of the corresponding citing papers. Furthermore, in Chapter 6 we consider author graph normalisation that also takes the number of authors on citing papers into consideration for normalisation. It should also be mentioned that the additional impact gained by papers through indirect citations for the Abramo method (Giuffrida *et al.*, 2018) is field normalised.

1.5 Test data

To evaluate the validity and utility of citation metrics different approaches exist. Axioms may be defined that describe desirable properties for well-defined metrics. Based on these axioms, metrics may be compared or validated for mathematical soundness under varying conditions (Altman and Tennenholtz, 2010; Bouyssou and Marchant, 2014, 2016). However, to empirically answer questions about metrics, an appropriate real-world publication data set is required. Frequently, correlation analyses between two or more metrics are performed. This can yield some insight into the scoring behaviour of the metrics but correlation comparisons are problematic on their own (West *et al.*, 2010; Thelwall, 2016). Furthermore, they only provide comparisons to some baselines, usually citation counts used as proxy for quality or impact.

An alternative approach is to use test data. One problem of evaluating metrics that measure academic quality or impact is the difficulty of obtaining appropriate test data. This problem is compounded by the subjectivity of what is considered quality or impact. Ideally one would collect direct peer-assessed test data through large-scale surveys. However, this is often rendered impractical since it is expensive, time-consuming, and difficult to obtain representative, unbiased, and sufficiently large enough test data sets. Bornmann and Marx (2015), for example, use a test data set of over 50 000 records obtained from a post-publication peer-review system of the biomedical literature. However, large-scale test data is openly available in very few cases.

In this dissertation we follow a different approach. We use a proxy for this assessment which comprises test data based on other ground truth provided by domain experts. The general assumption is that the test data entities exhibit some property that is not exclusively based on citations. Therefore, these test data sets are used to evaluate the functionality of a metric to identify the comprising entities and consequently their underlying shared property.

Recently, a few studies have used such relatively small test data sets to evaluate metrics in different application scenarios: to evaluate author metrics in identifying well-established researchers using test data that comprises researchers that have received fellowship status at learned societies, have won life-time contribution awards, or are frequently board members of prestigious journals (Fiala *et al.*, 2008; Nykl *et al.*, 2014; Fiala *et al.*, 2015; Gao *et al.*, 2016; Fiala and Tutoky, 2017); to evaluate paper-level metrics in finding impactful papers using test data that comprises high-impact paper awards (Sidiropoulos and Manolopoulos, 2005; Dunaiski and Visser, 2012; Mariani *et al.*, 2016); and to validate the applicability of newly proposed indicators (Gao *et al.*, 2016).

We collected five different test data sets and matched the comprising entities to the corresponding entities in the publication databases:

1. The **high-impact** paper test data set is a collection of 563 papers published between 1966 and 2014 that have won high-impact awards from conferences or organisations. These awards are handed out post-publication, usually 10–25 years after their initial publication, by selection committees comprising reviewers that can be assumed to be experts in the corresponding fields. Papers are typically evaluated on their continued impact in their field in terms of research, methodology, or application. For these high-impact papers the assumption is that they have had a long-lasting and influential impact on future papers. Therefore, we expect high-impact papers to have above average citation rates but also to have a latent property that is not encoded through pure citations.

2. The second test data set comprises 1119 papers published between 1962 and 2017 that have won **best paper awards** at different conferences or journals. The ratings are usually based on papers' intrinsic quality judged by reviewers and final decisions are made by editors or conference committees. Best paper awards are decided before or shortly after publication and therefore no or limited knowledge about future citation counts is available. For these best papers the underlying property is that they are high quality but might not have high impact. There are many other factors that influence the decisions of awarding best paper prizes that are not measurable through citations. Therefore, their underlying property is further detached from citation counts compared to the high-impact papers.

3. The **important papers** comprise a collection of 129 computer science papers considered important for the development of new fields within the computer science discipline. The source for this list is Wikipedia (Wikipedia, 2014) where papers that are regarded important to a research field are 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: (1) a publication led to a significant, new avenue of research in the domain in which it was published, (2) a paper is regarded as a breakthrough publication if it changed the scientific knowledge significantly, and (3) influential papers that had a substantial impact on the teaching of the domain. For this set of important papers the assumed underlying property is similar to the one of the high-impact paper test data set. However, we use this list cautiously since it is relatively small, may contain biases, and only contains computer science papers

and books. Therefore, results based on this test data set should not be generalised to all fields.

4. The test data set of **ACM fellows** comprises 1000 researchers that received an ACM fellowship between the years 1994 and 2015 in recognition of an individual's lasting impact on a field in computer science in terms of technical and leadership contributions, has influenced the direction of a field, and has to be evidenced by publications, awards, or other publicly recognised artefacts of merit. For the ACM fellows the underlying property is that they have had significant impact on a research field as demonstrated by papers and citation counts. However, we also assume that they have had some additional influence that is not only based on pure citation counts.

5. The last test data set comprises 596 researchers that have won achievement or **lifetime contribution awards** between the years 1958 and 2017. We considered awards that are handed out by conferences, learned societies, or special interest groups from different academic disciplines. Generally, the nomination processes consist of peer nominations and final decisions are taken by dedicated award committees. The underlying property of the researchers that have won lifetime contribution awards is similar to the ACM fellows.

We use different combinations of these test data sets for each chapter depending on the aim of the accompanying papers.

1.6 Evaluation measures

We use the terms metric, indicator, and ranking algorithm synonymously since they assign scores to academic entities that can be converted into a ranking (sorted list of entities with ascending ranks). When using test data (a subset of all entities considered relevant) to evaluate a ranking, some evaluation measure is required to translate the rank distribution of the relevant entities into a scalar-value performance score. In other words, metrics are used to rank entities, while evaluation measures are used to compute the performance of metrics.

Given test data comprising entities we assume to be relevant, we want to evaluate a metric's performance based on the ranks it assigns to them. Assume there are n relevant entities in a perfect test data set. An ideal metric would rank these entities in the top n ranks. However, in the context of citation analysis, the difficulty with ranked lists of real-world data is that they are usually orders of magnitude larger than the test data sets. Furthermore, the relevant entities are spread out substantially and do not necessarily have very high ranks.

The most elementary way to measure metrics' ranking performances is to compute the average or median rank of all relevant entities. Intuitively it makes sense that if one metric ranks the relevant entities higher on average than another metric, the former metric should be regarded as better than the latter. However, the average rank can easily be dominated by a small number of outliers in a skewed rank distribution. Furthermore, it is easy to attribute too much significance to a small change in the mean which ultimately might not be proportional to the difference in rankings and was simply observed by chance.

Many alternative evaluation measures exists, especially in information retrieval settings (for an overview see Chapter 8 in Manning *et al.* (2008, pp. 151–163)). However, many of these measures have not been considered for evaluating citation metrics. The most frequently used measures are the average, median, or sum of ranks (Fiala *et al.*, 2008; Fiala, 2012; Nykl *et al.*, 2014; Fiala *et al.*, 2015; Dunaiski *et al.*, 2016), the average ranking ratio (Nykl *et al.*, 2015; Mariani *et al.*, 2016), and the recall measure (Liu *et al.*, 2005; Dunaiski and Visser, 2012; Mariani *et al.*, 2016). Sometimes no evaluation measures are used and complete lists of ranks or scores are reported (Sidiropoulos and Manolopoulos, 2005; Gao *et al.*, 2016). Recently, alternative evaluation measures have been used such as the mean average precision measure (Dunaiski *et al.*, 2016) or the normalised discounted cumulative gain (nDCG) measure (Fiala and Tutoky, 2017).

In the setting of information retrieval problems, evaluation measures have been evaluated on their stability and sensitivity (Buckley and Voorhees, 2000; Voorhees and Buckley, 2002; Sakai, 2006). Sensitivity (or discriminative power) refers to a measure's ability to identify significant differences between sets of rankings, while stability refers to a metric's consistency in reporting the correct results under changing conditions. In the paper that accompanies Chapter 3, we describe the most common evaluation measures in the context of academic rankings and evaluate their fitness (stability and sensitivity) for measuring the performance of citation impact metrics.

Chapter 2

Evaluating indirect impact metrics using test data

In the paper which accompanies this chapter, we evaluate paper and author metrics using test data for the field of computer science. It was an initial effort to evaluate indirect metrics (various PageRank variants) on their ranking performances in comparison to citation counts. We use four different types of test data: (1) papers who have won high-impact awards for their continued influence in their computer science fields, (2) authors who have won contribution awards for significant contributions, (3) a list of papers, sourced from Wikipedia, claimed to be influential to computer science, and (4) a set of papers which were recognised as best paper at conferences or journals in a certain year. We reverse the use of the fourth test data set: instead of measuring the ranking performance of metrics, we assume that citation counts is an appropriate impact metric and evaluate how well the reviewers of the conferences and journals predict future highly-cited papers.

On the paper level we found that using citation counts is the best metric for ranking high-impact papers in general. However, when considering the list of influential papers, PageRank performs better than citation counts. On the author level we found that PageRank on the author citation graph performs the best in identifying the well-established authors, followed by citation counts without self-citations. Furthermore, we found that the g-index (Egghe, 2006) performs better than the h-index (Hirsch, 2005).

This was the first paper in which indirect paper-level metrics were evaluated and compared to direct citation metrics using larger test data sets. Previous evaluations only used a small number of data points and small subsets of publication databases (Sidiropoulos and Manolopoulos, 2005). Furthermore, we also analysed the impact that the time controlling parameters of PageRank-like algorithms have on the performance results and optimised these parameters using an appropriate evaluation methodology to avoid overfitting the test data by splitting it into stratified training, validation, and test sets. Lastly, we also showed that choosing appropriate citation window sizes is important since it impacts the results of the metrics.

The paper for this chapter:
Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, 10(2), 392–407.
Available at: http://dx.doi.org/10.1016/j.joi.2016.01.010

Chapter 3

How to evaluate rankings of academic entities using test data

In this paper we look at various aspects that are important when test data is used to evaluate metrics. We first discuss the prerequisite steps that are crucial when the aim is to obtain fair comparisons between different metrics. For example, we argue that scores produced by metrics should be converted to fractional ranks. However, we also point out that when using fractional ranks and evaluation measures based on precision, inconsistent results may be obtained where precision scores take on values larger than 1. This occurs when a test entity at position n has a fractional rank value that is smaller than n but can be addressed by simply setting an upper bound of 1. We also describe the most common evaluation measures, typically used in information retrieval settings, in the context of academic rankings and show how they can be adjusted for permille rankings.

With this paper, we transfer methodologies and best practices from the field of information retrieval to informetrics. The main work in this paper, however, is a second-order evaluation of the evaluation measures themselves. Buckley and Voorhees (2000) proposed a framework to compute the stability and sensitivity of evaluation measures (precision and recall at various cut-offs, and average precision) in information retrieval problems. Extensions of this methodology (Voorhees and Buckley, 2002; Sakai, 2006) can be used to estimate the minimum performance difference required by an evaluation measure to consider two rankings significantly different.

We adapt these methodologies and define a framework for rankings of academic entities where the rank distributions of test data are typically skewed and very sparse. With this framework we can now answer the question of which evaluation measure should be used for an experiment where different bibliometric metrics are compared. Furthermore, once an evaluation measure for an experiment is identified, we can now compute significance values associated with the performance differences between the metrics in the experiment.

For this paper we used two different publication databases and four test data sets to analyse the stability and discriminative power of various evaluation measures. We also demonstrate the functionality of the proposed framework through a set of 64 author and 38 paper ranking metrics (described in Dunaiski *et al.* (2016); Nykl *et al.* (2015); Dunaiski *et al.* (2018*a*)). For example, by randomly drawing different numbers of ranking metrics, we demonstrate that the sensitivity and stability of the evaluation measures remain relatively constant.

Using this framework to analyse the evaluation measures, we found that no clear winner exists and that their performance is highly dependent on the underlying data (test data and the database). We found that simple measures such as the average or median rank have high discriminative power and are stable evaluation measures. However, we also showed that relatively large performance differences are required to confidently determine if one ranking metric is significantly better than another. Lastly, we listed alternative measures that also yield stable results and highlighted measures that should not be used in the context of academic rankings. For example, we showed that the nDCG (Järvelin and Kekäläinen, 2002) measure is only appropriate when permille rankings are used.

The paper for this chapter:
Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). How to evaluate rankings of academic entities using test data. *Journal of Informetrics*, 12(3), 631–655.
Available at: https://doi.org/10.1016/j.joi.2018.06.002

The accompanying data for this chapter:
Dunaiski, M. (2018). Data for: How to evaluate rankings of academic entities using test data, Mendeley Data, v1. https://dx.doi.org/10.17632/4d46zncg4k.1.

The code for the associated tool is available at:
https://github.com/marceldunaiski/RankingEvaluation

Chapter 4

Ranking papers

In this paper we turn to the topic of normalisation of paper-level metrics. In addition to evaluating the metrics in terms of performance, we also quantify their ranking bias in terms of fields and time. The fairness test quantifies a metric's bias by sampling the top k percent of papers, counting to which fields they belong, and comparing this distribution to the field distribution of all papers in the used database (Radicchi and Castellano, 2012). The top k percent of papers ranked by a fair metric should have almost the same field distribution as the overall sample.

In this paper we use a methodology proposed by Vaccario *et al.* (2017) to evaluate metrics' field and time bias. It is based on the fairness test but includes two improvements. To simulate an unbiased ranking metric, a number of sampling processes are carried out in which k percent of papers are randomly drawn while recording their field distributions. These simulated distributions are then compared to the top k percent of papers and their field distribution ranked by an actual metric. This allows us to not only compute the per-field bias of metrics but also the proportion of a per-field bias to the overall bias of a metric. Furthermore, through the simulation of the random sampling process, confidence intervals may be computed characterising unbiased rankings.

In terms of performance evaluation, we use test data that consists of (1) papers that have won high-impact awards and (2) papers that have won prizes for outstanding quality. We consider different direct citation impact metrics and indirect ranking algorithms (PageRank variants and the Abramo method (Giuffrida *et al.*, 2018)) in combination with various normalisation approaches (mean-based, percentile-based, co-citation-based, and post hoc rescaling (Mariani *et al.*, 2016)). We conduct all experiments on the MAG and the ACM publication databases which use different field categorisations. On the MAG database, we use the top-level fields from its field categorisation scheme which is based on papers' semantic information. On the ACM database, we use the author-chosen concept categories of papers. When evaluating the metrics in terms of performance, we not only consider the age of papers through different citation window sizes but also the intrinsic ranking characteristics of the metrics. We show that some insight is gained through this which would have otherwise been missed.

We found that PageRank is less field biased, while citation counts are less time biased. This does not change when both metrics are normalised over fields and time. Furthermore, we found that PageRank's time bias is reduced when it is personalised with citation counts. However, its field bias is lower when no personalisation is used. Lastly, we also found that PageRank's damping factor has a large influence on its time bias but that it hardly impacts its field bias. When comparing percentile citation scores to mean-normalised citation scores, we found that the percentile approach is less field and time biased. However, we found no significant performance difference between these two metrics.

In terms of performance, we found that time-normalised metrics identify high-impact papers better shortly after their publication compared to their non-normalised variants. However, after 5 to 10 years, the non-normalised metrics perform better. A similar trend exists for the set of high-quality papers where these performance cross-over points occur after 5 to 10 years. We also found that PageRank should always be personalised with papers' citation counts and time-rescaled for citation windows smaller than 7 to 10 years.

The paper accompanying this chapter:
Dunaiski, M., Geldenhuys, J., & Visser, W. (2019). On the interplay between normalisation, bias, and performance of paper impact metrics. *Journal of Informetrics*, 13(1), 270–290.
Available at: https://doi.org/10.1016/j.joi.2019.01.003

The accompanying data for this chapter:
Dunaiski, M. (2018). "Data for: On the interplay between normalisation, bias, and performance of paper impact metrics", Mendeley Data, v1. http://dx.doi.org/10.17632/v4mxr9p3h5.1

Chapter 5

Ranking authors (part 1)

Two different approaches exist to aggregate normalised paper scores to higher-level size-independent impact metrics: the averaged and the globalised approach. For an aggregated research unit, the former approach computes the mean ratio of its actual and expected paper scores, while the latter computes the ratio of the sum of its actual paper scores and the sum of its expected paper scores. Opinions about which approach is more appropriate for research evaluation differ. Some argue for the averaged approach (Lundberg, 2007; Waltman *et al.*, 2011*a*), while others argue for the globalised approach (Moed, 2010; Vinkler, 2012).

In the paper accompanying this chapter, we investigate the differences between these two approaches. We use different paper-level impact metrics which use different normalisation approaches (mean-based, percentile-based, co-citation-based) and aggregate them to the author level. We then evaluate the two variants of each metric on field bias and performance on the author level. We found that the overall field bias between variants is very similar. In terms of performance, we found that metrics either perform better with the globalised approach or the difference is insignificant. For example, for paper scores based on citation counts or PageRank scores, the differences between the two variants are insignificant.

We also analysed the differences between the variants for a range of different citation window sizes. We highlight some bias and ranking trends that would otherwise have not been identified. We also considered the size-dependent variant of each metric, where authors' scores are the sum of their papers' normalised impact scores. The direct comparison of these variants showed that the RCR metric (Hutchins *et al.*, 2016) best identifies the well-established researchers early. For larger citation windows, we found that PageRank performs the best. Of the size-independent variants, RCR again performs the best for smaller citation windows. However for larger citation windows, PageRank performs the best on the multi-disciplinary database (MAG) while the Abramo method (Giuffrida *et al.*, 2018) performs the best on the computer science database (ACM).

The paper accompanying this chapter:
Dunaiski, M., Geldenhuys, J., & Visser, W. (2019). Globalised vs. averaged: Bias and ranking performance on the author level. *Journal of Informetrics*, 13(1), 299–313.
Available at: https://doi.org/10.1016/j.joi.2019.01.006

The data for the experiments in this paper is based on the data from the next chapter (Dunaiski, 2018).

Chapter 6

Ranking authors (part 2)

We reproduce and extend the work by Nykl *et al.* (2014) and Nykl *et al.* (2015) who evaluated various PageRank variations by analysing the effects that author graph normalisations, self-citations, and counting methods have on author rankings. Nykl *et al.* (2014) used 54 authors that have won one of two prestigious computer science awards, a set of 576 researchers that have received fellowships of the ACM, and a list of 280 highly cited researchers as test data. They used a subset of the Web Of Science (Clarivate Analytics, 2017) database for their experiments comprising 149 347 papers published in 386 computer science journals between 1996 and 2005. They found that the overall best PageRank approach to ranking well-established researchers is to compute PageRank on the paper citation graph personalised with author counts, to remove all self-citations, and to evenly distribute paper scores among co-authors.

Nykl *et al.* (2015) again use the ACM fellows as test data and two different lists of authors in the computer science fields ‘artificial intelligence’ and ‘hardware’ with 354 and 158 authors, respectively. These lists comprise authors that have won contribution awards, but also authors that have written papers that have won best paper awards or high-impact paper awards. They found that the overall best approach to rank high-impact authors is to use a paper’s journal impact score for personalisation of the paper-level PageRank computations.

With the paper that accompanies this chapter, we identify results that generalise by using larger test data sets and two publication databases, one of which is multidisciplinary. Furthermore, we use the methodology proposed in Dunaiski *et al.* (2018*b*) (Chapter 3) to ground these experiments methodologically on a more solid foundation. Lastly, we include additional author impact metrics in the evaluation: the percentile-based R6 metric (Leydesdorff *et al.*, 2011) and the PR-index (Gao *et al.*, 2016), which combines PageRank and a variant of the h-index (Hirsch, 2005).

We describe here briefly some of the results that seem to be generally true and that do not depend on the databases used. PageRank proves to be the best metric for ranking the well-established authors and outperforms the R6 metric, as well as the more traditional impact metrics such as citation counts and the h-index. We also show that it is more important to personalise PageRank appropriately on the paper level than deciding whether to include or exclude self-citations. In general, the best results are obtained when PageRank is personalised with papers’ corresponding journal impact values. However, on the author level, we find that author graph normalisation is more important than personalisation.

Self-citations play an important role for all metrics. We found that the improvements by only removing direct author self-citation on the author graph are not significant for most metrics. However, when removing all co-author self-citations the performances of all metrics are significantly improved. When considering different counting methods, we found that evenly distributing a paper's score between co-authors always yields the best results for ranking well-established researchers, irrespective of what paper impact indicator is used. Lastly, we also found that the PR-index is better in identifying well-established researchers compared to the more commonly used h-index and g-index.

From the experiments and results in this paper, we made two important observations: (1) evenly sharing paper credit among co-authors is generally the best approach and does not require knowledge about the author positions on papers and (2) computing PageRank on the author citation graph is computationally much more expensive than computing PageRank directly on the paper citation graph and does not significantly improve the results. These two findings suggest that PageRank on the paper citation graph is the recommended approach for computing author impact scores of well-established researchers.

The paper for this chapter:
Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). Author ranking evaluation at scale. *Journal of Informetrics*, 12(3), 679–702.
Available at: https://doi.org/10.1016/j.joi.2018.06.004

The data for the experiments in this paper:
Dunaiski, M. (2018). Data for: Author ranking evaluation at scale. Mendeley data, v1. https://dx.doi.org/10.17632/5tzchw6r6d.1.

Chapter 7

Discussion and conclusion

Validation with test data Using test data to evaluate and compare impact metrics is a valuable tool in scientometrics. However, it is important not to overstate or generalise findings based on small test data sets. In the work presented with this dissertation we found many occurrences where the outcome of experiments showed different results depending on the type of test data used and on which database the experiments were executed. It is therefore important that conclusions are restricted to the characteristics of the test data and not put forward as general claims. Furthermore, it shows that it is also important to repeat experiments on different data to validate findings. In general, we used one multi-disciplinary database and one that only contains a single discipline (computer science). We found that many results differed between databases, but some results seem to be database independent. More importantly, we found that data selection is crucial since it impacts the results substantially.

For robust analyses, it is very important that standardised (acknowledged), high-quality (unbiased), and openly available test data is collected. Post-publication rating systems similar to F1000Prime[1] would be highly valuable if they were openly available and multi-disciplinary. Either comprehensive cross-disciplinary test data sets are used, or studies have to be repeated for multiple disciplines until general truths are identified.

We spent a lot of time collecting and cleaning the test data by hand from various Internet resources. Matching papers to database entries was relatively trivial, especially where DOIs were available. However, cleaning and matching researchers' names to corresponding author entities in the databases were very time-consuming.[2] We described this procedure in the paper that accompanies Chapter 6. For the sake of reproducibility and the hope that researchers apply test data driven validation in future studies we have published all data that we collected. We hope that future scholars update and diversify the available test data, especially for fields that are underrepresented in our test data sets.

Only using a single citation window size will only report on the best metric for that specific citation window size. We showed in Chapters 4 and 5 that the relative performances between the various metrics change dramatically when citation window sizes are varied. Furthermore, it is also important to put performance results of metrics into context of their intrinsic ranking characteristics (Chapter 4). We showed this by contrasting the metrics' ranking trends using the actual test data and their ranking trends using multiple samples of random entities that are year-stratified according to the actual test

[1] https://f1000.com/prime/rankings

[2] We coded a graphical user interface to help streamline the matching, author name disambiguation, and entity merging processes for the MAG database.

data. The intrinsic ranking characteristic of a metric therefore reflects a metric's expected ranking trend for a set of entities with a certain age distribution within a citation graph.

This methodology also yields some insight into whether the test data comprises some latent property which differentiates the test entities from a set of "average" entities. For a potential future investigation, one could formally incorporate the notion of intrinsic ranking characteristics into evaluation measures. For example, assume that metrics are evaluated in terms of recall. The score of each entity may be weighted differently (usually 1), therefore changing their contributions to the overall evaluation score. The open question is how to weight the entities appropriately as a function of the expected ranking trends of metrics. Mariani *et al.* (2016) use a similar approach in which test entities are penalised if more than the expected number of entities belong to the same field. In other words, a metric is penalised if its rankings are biased towards certain fields.

A drawback of unlabelled test data, such as the data used in this dissertation, is that evaluations are based on ratings that are binary (i.e., an entity is either relevant or not). If instead a rating scale is used for the test entities, it would give more flexibility and variety in statistical tests. In addition, some evaluation measures such as nDCG (Järvelin and Kekäläinen, 2002) are defined for scale-graded test entities. Furthermore, the results would likely be more fine-grained and multifaceted. Considering the test data used in this dissertation, various options are available to create scaled ratings. For example, papers that have won high-impact awards may be assigned different grades corresponding to the prestigiousness of the awards. This can either be accomplished through surveys of researchers in the corresponding fields (Zheng and Liu, 2015) or by using some impact indicator for the associated conferences and journals. For the author test data one could grade the test authors, for example, based on how many awards they have received.

The merit of PageRank One of the obvious questions to ask is whether there is merit in using PageRank as an impact indicator for academic entities. To answer this question, one has to carefully determine what PageRank scores actually indicate. Does PageRank's output reflect impact or importance, and is it able to decode the rather latent attribute of quality? Martin (1996) contrasts between importance and impact where he describes the former as the potential influence of a paper on related research activities (i.e., future papers). He describes the latter as the actual impact on related research at a given time and argues that it is closer related to papers' citation counts.

On the journal level, PageRank has definite merit and has taken on the role as an important alternative measure to indicate a journal's value. Both Web of Science (WoS) and Scopus include journal scores based on the PageRank paradigm. WoS reports the Article Influence (AI) score for journals (Bergstrom, 2007; Bergstrom *et al.*, 2008) which is a size-independent version of PageRank on a journal cross-citation graph excluding journal self-citations. Scopus reports the SJR2 indicator (Guerrero-Bote and Moya-Anegón, 2012), a revised version of the SJR indicator (González-Pereira *et al.*, 2010), which adds a feature that citations from journals that are more closely related (based on co-citations) are weighted more than citations from distant journals.

Walters (2014) claims that the impact factor (direct citations) and AI (PageRank) both measure impact but does not support this claim with any credible arguments. Similarly, Davis (2008) argues that there is little difference between total citation counts and scores based on the Eigenfactor metric (the size-dependent variant of AI) due to high correlation values. However, this claim has been rebutted by West *et al.* (2010). The consensus seems to be that citation counts of journals measure popularity while scores based on PageRank indicate prestige (Bollen *et al.*, 2006; Franceschet, 2010; West *et al.*, 2010; Guerrero-Bote and Moya-Anegón, 2012).

However, in this dissertation we only analysed PageRank on the paper and the author levels where it is less often applied. Let us first consider the case of PageRank on the paper level, where its drawbacks have been discussed extensively. The biggest drawback is the fact that PageRank is biased towards older nodes in the citation network (Walker *et al.*, 2007; Chen *et al.*, 2007; Dunaiski, 2014; Mariani *et al.*, 2016; Vaccario *et al.*, 2017). In Chapter 2 we used computer science test data and databases comprising only computer science literature. Comparing PageRank and various variants to limit its time bias to standard citation counts, we found no clear advantage of using PageRank, even when no citation windows are specified and complete citation graphs are considered. However, in Chapter 4 using larger test data sets from more disciplines, we found that PageRank can outperform citation counts. In general, this occurs when larger citation windows are used.

Although we evaluated PageRank in terms of an impact metric with a test data set comprising high-impact papers, we argue that PageRank on the paper level indicates a paper's influence well, if influence is interpreted as the potential influence on future papers (Martin, 1996). We base this argument on the observation that PageRank performs especially well in identifying the high-impact papers when they are older. The high-impact papers in the test data set are chosen with citation counts in mind but other selection criteria also play a role such as their continued impact. It is reasonable to assume that these types of papers were selected because they still have current relevance. In the paper accompanying Chapter 4 we show that their rankings based on citation counts vary significantly from rankings based on PageRank. PageRank, on the MAG database, identifies these high-impact papers much better, even when considering the intrinsic ranking characteristic (time bias) of PageRank. Since PageRank identifies these papers better, especially for larger citation windows, we argue that PageRank is better in identifying papers with continued influence compared to only looking at citation counts.

In terms of PageRank's ability to identify the latent attribute of paper quality we tried to gain some insight by using test data that comprises papers that have won best paper awards. Drawing conclusions about results based on this test data is tricky since many unknown factors can influence the selection of papers that win these types of awards. However, we compared the ranking results of PageRank against citation counts based on this test data set and found similar results as for the high-impact papers. That is, on the multi-disciplinary database, PageRank does identify these papers better than citation counts for any citation window size (Dunaiski *et al.*, 2019). It does so, although according to their relative expected ranking performances, PageRank should only perform better than citation counts for larger citation windows. This holds true when both normalised and non-normalised variants are compared.

Therefore, on the paper level, PageRank's ideal application would be to identify influential papers with continued impact. We also found that PageRank performs better on the multi-disciplinary database compared to the database comprising only computer science papers. Furthermore, PageRank's algorithm is complex compared to commonly used impact metrics which makes the interpretation of PageRank scores more difficult, thereby limiting its application. Therefore, PageRank has limited value for small-scale, personal, and field-specific bibliometric analyses. However, for large-scale bibliometric analyses (i.e., application in search engines) it certainly has merit.

When units are authors, PageRank may be computed on two different networks. One may compute PageRank on the author co-citation graph where nodes are authors and edges are produced when a paper by a citing author cites a paper by a cited author (Ding *et al.*, 2009; West *et al.*, 2013). Alternatively, one may compute PageRank scores for papers directly on the paper citation graph and aggregate them to the author level. Both approaches yield size-dependent metrics. These two approaches may be converted to size-independent variants by dividing authors' PageRank scores by their number of papers. For the former approach the resulting score reflects the average influence/impact per paper of an author's papers, while the latter indicates an author's average influence/impact per paper. Instead of dividing authors' PageRank scores by their number of papers, one may compute their proportions of highly scored papers. This results in a measure of "high influence". However, the soundness and merit of these variants have to be investigated in future studies.

Furthermore, future studies should also investigate how PageRank violates consistency axioms. For example, it is unclear in which circumstances metrics based on PageRank violate the property of size-dependent metrics that adding citations and papers reduce the rank of papers and authors, respectively. Furthermore, an open question is whether situations can occur in which nodes change rank orders when they increase their in-degree by the same amount. Irrespective of PageRank's violation of these axioms, we found that it has some ranking advantages. For example on the author level, PageRank has some distinct advantages over citation counts. We showed that it clearly performs better in ranking well-established researchers for larger citation windows. On a multi-disciplinary database it also shows substantially reduced field bias. Furthermore, without normalisation steps, PageRank shows less topical and more temporal bias than citation counts. With additional normalisation, PageRank improves over algorithms that already incorporate normalisation. When using PageRank to rank well-established researchers we found that it is important to consider the venues at which they publish their work. We also found that normalisation of the underlying author co-citation graph has a significant impact on PageRank.

New data sources should be exploited to fine-tune PageRank. Instead of defining new PageRank variants, we recommend that the basic PageRank algorithm be used but with altered parameters. Various possibilities exists that may or may not make use of new data sources:

1. The underlying graph may be updated by adding or removing appropriate citations. This might be especially important for the paper citation graph since its time bias may be reduced by adding appropriate forward citations. For example, by using topics or semantic similarities between papers, forward references may be added to similar papers thereby creating feedback loops.

2. Information from new data sources may be used to rescale PageRank's personalisation vector appropriately. For example, papers' read or download counts may be used as a popularity estimate to obtain a better simulation of real-world searches.

3. Mariani *et al.* (2016) compare time-rescaled PageRank to CiteRank (Walker *et al.*, 2007) which incorporates a time decay factor for older papers. Mariani *et al.* (2016) found that time-rescaling standard PageRank shows less bias and better ranking performance in terms of ranking important papers in the field of physics. Based on these results, we suggest that PageRank should be time-rescaled post hoc (Vaccario *et al.*, 2017) as we did in Chapters 4 and 5 instead of trying to reduce PageRank's time bias by introducing time-decay factors or changing edge weights in the citation graph.

4. The increasing adoption of more specific author contribution statements on papers gives many opportunities for inclusion into PageRank's personalisation vector, if computed on the author cross-citation graph. However, future investigations would have to identify whether this may improve the author-level PageRank performance over the paper-level performance before aggregation.

To summarise while exercising caution not to overvalue the identified trends that lead to misleading generalisations, below are some additional observations we found as a result of our investigation:

- In terms of ranking well-established authors and without restricting citation windows, we found that PageRank performs the best when paper-level self-citations are removed and paper scores are evenly shared among co-authors (fractional counting). This confirms the results found by Nykl *et al.* (2014). Furthermore, we found that the best personalisation approach is to use papers' journal impact scores on the paper graph, confirming the results of Nykl *et al.* (2015). In general, we found that either no personalisation or per-paper (size-independent) journal influence values yield the best results.

- On the author cross-citation graph, Fiala *et al.* (2015) found that PageRank-like impact metrics yield no significant improvement over using citation counts. In Chapter 6, we found that using PageRank, either on the paper citation graph or on the author citation graph, significantly outperforms citation counts. However, in Chapter 5, we showed that these generalised results only show part of the whole picture. In fact, when comparing metrics on their relative ranking performances, citation windows are highly important to consider since results will change depending on the citation window size used. In Chapter 5, we found that PageRank does perform better than citation counts, but only for larger citation windows. For the size-dependent variants of PageRank and citation counts, we found that PageRank only performs better after 18 years on the ACM database. On the MAG database, it always performs better. However, it should be noted that these results are based on field- and time-normalised scores.

- Without normalisation paper-level PageRank shows less topical and more temporal bias compared to citation counts. This supports the results found by Vaccario *et al.* (2017). When PageRank is time and field normalised it has less overall bias compared to time and field normalised citation counts. Furthermore, on the author level, when PageRank is field and year normalised its temporal bias is significantly reduced but its field bias remains about the same (Chapter 5).

The RCR metric There is certainly some merit in the approach of the RCR metric, although its specific formulation has been criticised for various reasons. Bornmann and Haunschild (2017) found that RCR has a relatively low correlation with peer-rated papers based on the F1000Prime data. However, other citing- and cited-side normalised indicators have equally low correlations. Furthermore, they found that RCR has relatively high correlation with these metrics, which is not surprising since they are all based on citation counts. Janssens *et al.* (2017) criticise RCR's time normalisation approach since it uses a paper's citation count divided by the number of years since publication. They argue that through this rescaling approach, RCR scores will decline for older papers since papers in general tend to attract fewer citations with age. Therefore, the RCR score reflects a paper's average influence over time and not the influence it had when it had the most impact and was frequently cited. On the author level, this depreciation of older papers may unfairly disadvantage established and mid-career researchers since their average RCR scores are negatively effected by the decreased RCR scores of their older papers. Hutchins *et al.* (2017) show that the majority of papers retain their influence over a researcher's career and claim that RCR does not unfairly disadvantage older papers. Furthermore, they argue that this property of the RCR is necessary and important, since some work will inevitably lose influence when it is replaced with newer work. Lastly, it should be mentioned that the RCR scores are unstable for papers published less than two to three years ago (Hutchins *et al.*, 2016) and for papers with fewer than five citations (Hutchins *et al.*, 2017). Furthermore, Janssens *et al.* (2017) argue that these thresholds should be chosen even more conservatively.

We did not analyse whether RCR paper scores are stable over time. We only analysed RCR's ranking trends based on our test data sets with different citation window sizes (Chapter 4). We found that RCR does have a substantial decline in ranking performance for larger citation windows, which may be explained by the above mentioned score depreciation for older papers. RCR does rank the high-impact papers higher than the other considered metrics for the first two or three years, which are citation window sizes for which it is known to be unstable. Furthermore, when considering RCR's intrinsic ranking trends, it does not perform above expectation compared to the other metrics, since it naturally ranks papers higher for smaller citation windows. When the intrinsic ranking trends are considered we found that other metrics actually perform above expectation. Specifically, on the ACM database, time-normalised citation counts perform better than RCR for $t < 6$ after which standard citation counts perform better than both. On the MAG database, time-normalised PageRank performs better for $5 < t < 9$, after which standard PageRank performs better than both. For citation windows smaller than five years, it remains to be investigated whether RCR actually identifies high-impact papers better or whether this performance is caused by its natural ranking trend compared to other metrics.

On the author level, we found that RCR performs well in identifying well-established authors early (Chapter 5). However, it is unclear whether the better performance of the RCR metric is due to using co-cited papers as reference sets or whether it is because the reference sets are normalised by the citation scores of the journals at which the papers are published. Furthermore, whether this is due to RCR's time normalisation approach remains an open question.

Janssens *et al.* (2017) also criticise the use of co-cited papers to specify papers' fields used as reference sets for score normalisation. They found that most of the co-cited papers, that are in a paper's co-citation network due to only a few co-citations, do not belong to the same topic. They found that 80% of papers are co-cited only once. However, restricting a paper's field to papers that are co-cited at least twice already reduces the reference set sizes to such an extent that it would generate unstable RCR values, even for articles with a reasonable number of citations.

On the paper level, we found that the RCR metric has relatively low field bias compared to standard citation counts. On the computer science database, RCR also has less field bias compared to PageRank. However, on the multi-disciplinary database, PageRank has substantially less field bias, presumably since PageRank incorporates citing-side citation normalisation. Lastly, in terms of time bias we found that only standard citation counts and PageRank exhibit more bias than RCR, which may be explained by the age normalisation of paper scores in the RCR metric. On the author level we found that RCR as a size-dependent metric has less field bias than the other considered metrics for all citation window sizes up to 25 years. As a size-independent metric, we found that RCR has less field bias than citation counts on the multi-disciplinary database, but that it has more bias on the computer science database.

Evaluation measures In Chapter 3 we analysed the sensitivity and stability of evaluation measures in the context of rankings produced by bibliometric impact metrics. We focused on evaluation measures typically used in information retrieval problems. However, some additional measures should be evaluated in future studies, such as the average ranking ratio (Mariani *et al.*, 2016) and performance scores based on computing the sum of test entities' relative ranks (Nykl *et al.*, 2014).

The average ranking ratio requires multiple rankings that are compared for its computation. Specifically, the rank of an entity from the metric that is evaluated is compared to the entity's ranks from all other metrics under consideration. (See Equation B.1 in Appendix B in Dunaiski *et al.* (2019) for the definition.) In order to measure its sensitivity and stability for comparisons to other evaluation measures, the framework discussed in Chapter 3 would have to be adapted trivially. Instead of a query only considering the ranks of a single metric, each query also requires the rankings of all other metrics under consideration. More difficult are the generalisations of the results since they are even more dependent on the set of rankings that are compared. In other words, the performance of the average ranking ratio might vary significantly between different experiments. For individual studies, we recommend that the average ranking ratio be considered as a potential evaluation measure, since it dampens the effects of outliers and uses per-entity comparisons instead of comparing summary scores from rank distributions.

The list below briefly summarises our contributions and findings relating to evaluation measures in the context of bibliometric impact metrics:

- Only reporting the scores obtained from an evaluation measure has two notable insufficiencies. Firstly, one may question whether the applied evaluation measure is appropriate given the rank distribution of the test data. Secondly, the significance of the observed performance differences when metrics are compared remains unknown. We addressed these two shortcomings in Chapter 3.

- We showed that the performance of evaluation measures, in terms of sensitivity and stability, on skewed and sparse rank distributions is highly dependent on the underlying data. This signifies that selecting an appropriate evaluation measure is important. However, we showed that simple measures such as the average or median rank measure are generally stable and have good discriminative power. We also found that score differences in average and median ranks required to confidently differentiate rankings are surprisingly high.

- We found that Precision and nDCG perform better when the long tails (lower ranked entities) of rank distributions are ignored by specifying a cut-off threshold. We found that these two evaluation measures with cut-offs defined by the average rank consistently have high discriminative power. Furthermore, we showed that nDCG's performance improves when the evaluated rankings are converted from absolute ranks to permille ranks. However, for other evaluation measures there is no significant improvement when permille rankings are used.

- We argued that when cut-off thresholds are specified they should not be defined by static values or by the size of the test data set. Instead they should be based on the ranks of the relevant entities of the test data set. We found that specifying the cut-off threshold at the average rank of the relevant entities or the rank at which 50% recall is achieved, generally performs better. Further research is required to define cut-off thresholds more appropriately and to investigate how threshold variation impacts the performance of evaluation measures.

- Lastly, in the paper that accompanies Chapter 4 we showed how the evaluation framework (Chapter 3) can be adapted to evaluate measures that are based on multiple queries. We used this adapted framework to compare different multi-query evaluation measures based on average ranks. We compared the unweighted mean average rank (MAR) measure and a weighted mean average rank (MAR_w) measure to the standard average rank measure and found that they all perform very similarly.

Conclusion To evaluate impact metrics with test data conscientiously is a relatively difficult and intricate task. We found that results are very unstable and change quickly when parameters are slightly changed, test data is varied, different databases are used, and differently sized citation windows are specified. Furthermore, the outcome of performance results also change occasionally when different evaluation measures are used. Nonetheless, evaluating impact metrics with appropriate test data is a useful and necessary tool in scientometrics to better understand their ranking behaviours. Therefore, it is essential that best practices are identified and applied for these types of investigations. With the work presented in this dissertation, we tried to fill some of the gaps on this topic.

List of references

Abramo, G., Cicero, T. and D'Angelo, C.A. (2011). Assessing the varying level of impact measurement accuracy as a function of the citation window length. *Journal of Informetrics*, vol. 5, no. 4, pp. 659–667.

Abramo, G., Cicero, T. and D'Angelo, C.A. (2012). A sensitivity analysis of researchers' productivity rankings to the time of citation observation. *Journal of Informetrics*, vol. 6, no. 2, pp. 192–201.

ACM, Inc. (2017). The 2012 ACM Computing Classification System. http://www.acm.org/publications/class-2012-intro. [Online; accessed 12-Sep-2018].

Adams, J., Gurney, K. and Jackson, L. (2008). Calibrating the zoom – a test of Zitt's hypothesis. *Scientometrics*, vol. 75, no. 1, pp. 81–95.

Ajiferuke, I., Burell, Q. and Tague, J. (1988). Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, vol. 14, no. 5, pp. 421–433.

Aksnes, D.W. (2003). A macro study of self-citation. *Scientometrics*, vol. 56, no. 2, pp. 235–246.

Altman, A. and Tennenholtz, M. (2010). An axiomatic approach to personalized ranking systems. *Journal of the ACM*, vol. 57, no. 4, pp. 1–35.

Assimakis, N. and Adam, M. (2010). A new author's productivity index: P-index. *Scientometrics*, vol. 85, no. 2, pp. 415–427.

Bergstrom, C.T. (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *College and Research Libraries News*, vol. 68, no. 5, pp. 314–316.

Bergstrom, C.T., West, J.D. and Wiseman, M.A. (2008). The eigenfactor™ metrics. *Journal of Neuroscience*, vol. 28, no. 45, pp. 11433–11434.

Bollen, J., Rodriquez, M.A. and Van de Sompel, H. (2006). Journal status. *Scientometrics*, vol. 69, no. 3, pp. 669–687.

Bornmann, L. (2017). Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, vol. 73, no. 5, pp. 775–787.

Bornmann, L. and Haunschild, R. (2017). Relative Citation Ratio (RCR): An empirical attempt to study a new field-normalized bibliometric indicator. *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 1064–1067.

Bornmann, L., Leydesdorff, L. and Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, vol. 7, no. 1, pp. 158–165.

Bornmann, L. and Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, vol. 9, no. 2, pp. 408–418.

Bornmann, L. and Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, vol. 66, no. 11, pp. 2215–2222.

Bouyssou, D. and Marchant, T. (2014). An axiomatic approach to bibliometric rankings and indices. *Journal of Informetrics*, vol. 8, no. 3, pp. 449–477.

Bouyssou, D. and Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. *Journal of Informetrics*, vol. 10, no. 1, pp. 183–199.

Brin, S. and 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. Elsevier Science Publishers B. V., Amsterdam, The Netherlands.

Buckley, C. and Voorhees, E.M. (2000). Evaluating evaluation measure stability. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, pp. 33–40. ACM, New York, NY, USA.

Chen, P., Xie, H., Maslov, S. and Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15.

Clarivate Analytics (2017). Web of science. https://www.webofknowledge.com. [Online; accessed 07-Sep-2017].

Colliander, C. and Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, vol. 5, no. 1, pp. 101–113.

Costas, R., van Leeuwen, T.N. and Bordons, M. (2010). Self-citations at the meso and individual levels: effects of different calculation methods. *Scientometrics*, vol. 82, no. 3, pp. 517–537.

Costas, R., van Leeuwen, T.N. and van Raan, A.F.J. (2011). The "mendel syndrome" in science: durability of scientific literature and its effects on bibliometric analysis of individual scientists. *Scientometrics*, vol. 89, no. 1, pp. 177–205.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, vol. 52, no. 7, pp. 558–569.

Davis, P.M. (2008). Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts? *Journal of the American Society for Information Science and Technology*, vol. 59, no. 13, pp. 2186–2188.

Ding, Y., Yan, E., Frazho, A. and Caverlee, J. (2009). PageRank for ranking authors in co-citation networks. *Journal of the American Society for Information Science and Technology*, vol. 60, no. 11, pp. 2229–2243.

Dunaiski, M. (2014). *Analysing Ranking Algorithms and Publication Trends on Scholarly Citation Networks*. Master's thesis, Stellenbosch University.

Dunaiski, M. (2018). Data for: Author ranking evaluation at scale. Mendeley data, v1. http://dx.doi.org/10.17632/5tzchw6r6d.1.

Dunaiski, M., Geldenhuys, J. and Visser, W. (2018*a*). Author ranking evaluation at scale. *Journal of Informetrics*, vol. 12, no. 3, pp. 679–702.

Dunaiski, M., Geldenhuys, J. and Visser, W. (2018*b*). How to evaluate rankings of academic entities using test data. *Journal of Informetrics*, vol. 12, no. 3, pp. 631–655.

Dunaiski, M., Geldenhuys, J. and Visser, W. (2019). On the interplay between normalisation, bias, and performance of paper impact metrics. *Journal of Informetrics*, vol. 13, no. 1, pp. 270–290.

Dunaiski, M. and 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. ACM, New York, NY, USA.

Dunaiski, M., Visser, W. and Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, vol. 10, no. 2, pp. 392–407.

Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics*, vol. 69, no. 1, pp. 131–152.

Egghe, L. and Rousseau, R. (1996). Averaging and globalising quotients of informetric and scientometric data. *Journal of Information Science*, vol. 22, no. 3, pp. 165–170.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, vol. 6, no. 3, pp. 370–388.

Fiala, D., Rousselot, F. and Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, vol. 76, no. 1, pp. 135–158.

Fiala, D., Šubelj, L., Žitnik, S. and Bajec, M. (2015). Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, vol. 9, no. 2, pp. 334–348.

Fiala, D. and Tutoky, G. (2017). PageRank-based prediction of award-winning researchers and the impact of citations. *Journal of Informetrics*, vol. 11, no. 4, pp. 1044–1068.

Fragkiadaki, E. and Evangelidis, G. (2014). Review of the indirect citations paradigm: theory and practice of the assessment of papers, authors and journals. *Scientometrics*, vol. 99, no. 2, pp. 261–288.

Franceschet, M. (2010). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, vol. 4, no. 1, pp. 55–63.

Gao, C., Wang, Z., Li, X., Zhang, Z. and Zeng, W. (2016). PR-Index: Using the h-Index and PageRank for Determining True Impact. *Plos One*, vol. 11, no. 9, p. e0161755.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, vol. 178, no. 4060, pp. 471–479.

Garfield, E. (1979). Is Citation Analysis A Legitimate Evaluation Tool? *Scientometrics*, vol. 1, no. 4, pp. 359–375.

Gazni, A., Sugimoto, C.R. and Didegah, F. (2011). Mapping world scientific collaboration: Authors, institutions, and countries. *Journal of the American Society for Information Science and Technology*, vol. 63, no. 2, pp. 323–335.

Giuffrida, C., Abramo, G. and D'Angelo, C.A. (2018). Do all citations value the same? Valuing citations by the value of the citing items. `https://arxiv.org/abs/1809.06088`.

González-Pereira, B., Guerrero-Bote, V.P. and Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, vol. 4, no. 3, pp. 379–391.

Guerrero-Bote, V.P. and Moya-Anegón, F. (2012). A further step forward in measuring journals' scientific prestige: The SJR2 indicator. *Journal of Informetrics*, vol. 6, no. 4, pp. 674–688.

Harzing, A.-W. and Alakangas, S. (2017). Microsoft Academic: is the phoenix getting wings? *Scientometrics*, vol. 110, no. 1, pp. 371–383.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16569–16572.

Howard, G.S., Cole, D.A. and Maxwell, S.E. (1987). Research productivity in psychology based on publication in the journals of the American Psychological Association. *American Psychologist*, vol. 42, no. 11, pp. 975–986.

Hug, S.E. and Brändle, M.P. (2017). The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics*, vol. 113, no. 3, pp. 1551–1571.

Hug, S.E., Ochsner, M. and Brändle, M.P. (2017). Citation analysis with Microsoft Academic. *Scientometrics*, vol. 111, no. 1, pp. 371–378.

Hutchins, B., Hoppe, T., Meseroll, R., Anderson, J. and Santangelo, G. (2017). Additional support for RCR: A validated article-level measure of scientific influence. *PLoS Biology*, vol. 15, no. 10, p. e2003552.

Hutchins, B.I., Yuan, X., Anderson, J.M. and Santangelo, G.M. (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biology*, vol. 14, no. 9, p. e1002541.

Hwang, W., Chae, S., Kim, S. and 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. ACM, New York, USA.

Ioannidis, J.P.A., Boyack, K. and Wouters, P.F. (2016). Citation Metrics: A Primer on How (Not) to Normalize. *PLoS Biology*, vol. 14, no. 9, p. e1002542.

Janssens, A., Goodman, M., Powell, K. and Gwinn, M. (2017). A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). *PLoS Biology*, vol. 15, no. 10, p. e2002536.

Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446.

Lange, L.L. (2001). Citation counts of multi-authored papers – First-named authors and further authors. *Scientometrics*, vol. 52, no. 3, pp. 457–470.

Larivière, V. and Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, vol. 5, no. 3, pp. 392–399.

Larivière, V., Gingras, Y., Sugimoto, C.R. and Tsou, A. (2014). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology*, vol. 66, no. 7, pp. 1323–1332.

Leydesdorff, L., Bornmann, L., Mutz, R. and Opthof, T. (2011). Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents. *Journal of the American Society for Information Science and Technology*, vol. 62, no. 7, pp. 1370–1381.

Lindsey, D. (1980). Production and citation measures in the sociology of science: The problem of multiple authorship. *Social Studies of Science*, vol. 10, no. 2, pp. 145–162.

Liu, X., Bollen, J., Nelson, M.L. and Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, vol. 41, no. 6, pp. 1462–1480.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics*, vol. 1, no. 2, pp. 145–154.

Manning, C.D., Raghavan, P. and Schütze, H. (2008). *Introduction to Information Retrieval.* Cambridge University Press, New York, NY, USA.

Mariani, M.S., Medo, M. and Zhang, Y.C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, vol. 10, no. 4, pp. 1207–1223.

Martin, B.R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, vol. 36, no. 3, pp. 343–362.

Meho, L.I. and Yang, K. (2007). Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the American Society for Information Science and Technology*, vol. 58, no. 13, pp. 2105–2125.

Microsoft (2017). Microsoft Academic Graph. `https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/`. [Online; accessed 15-Aug-2017].

Mingers, J. and Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, vol. 246, no. 1, pp. 1–19.

Moed, H.F. (2010). Cwts crown indicator measures citation impact of a research group's publication oeuvre. *Journal of Informetrics*, vol. 4, no. 3, pp. 436–438.

Nykl, M., Campr, M. and Ježek, K. (2015). Author ranking based on personalized PageRank. *Journal of Informetrics*, vol. 9, no. 4, pp. 777–799.

Nykl, M., Ježek, K., Fiala, D. and Dostal, M. (2014). PageRank variants in the evaluation of citation networks. *Journal of Informetrics*, vol. 8, no. 3, pp. 683–692.

Parolo, P.D.B., Pan, R.K., Ghosh, R., Huberman, B.A., Kaski, K. and Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, vol. 9, no. 4, pp. 734–745.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, vol. 12, no. 5, pp. 297–312.

Pudovkin, A. and Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. *Collnet Journal of Scientometrics and Information Management*, vol. 3, no. 2, pp. 3–10.

Radicchi, F. and Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, vol. 6, no. 1, pp. 121–130.

Radicchi, F., Fortunato, S. and Castellano, C. (2008). Universality of citation distributions: toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 45, pp. 17268–17272.

Rosvall, M. and Bergstrom, C.T. (2011). Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one*, vol. 6, no. 4, p. e18209.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. In: *Proceedings of the 29th Annual International ACM SIGIR conference on Research and development in information retrieval*, SIGIR '06, pp. 525–532. ACM, New York, NY, USA.

Sidiropoulos, A. and Manolopoulos, Y. (2005). A citation-based system to assist prize awarding. *ACM SIGMOD Record*, vol. 34, no. 4, pp. 54–60.

Silva, F.N., Rodrigues, F.A., Oliveira, O.N. and da F. Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, vol. 7, no. 2, pp. 469–477.

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, vol. 108, no. 1, pp. 337–347.

Trueba, F.J. and Guerrero, H. (2004). A robust formula to credit authors for their publications. *Scientometrics*, vol. 60, no. 2, pp. 181–204.

Vaccario, G., Medo, M., Wider, N. and Mariani, M.S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics*, vol. 11, no. 3, pp. 766–782.

van Leeuwen, T.N. and Calero Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation*, vol. 21, no. 1, pp. 61–70.

van Raan, A.F.J., van Leeuwen, T.N. and Visser, M.S. (2011). Severe language effect in university rankings: particularly germany and france are wronged in citation-based rankings. *Scientometrics*, vol. 88, no. 2, pp. 495–498.

Vinkler, P. (2012). The case of scientometricians with the "absolute relative" impact indicator. *Journal of Informetrics*, vol. 6, no. 2, pp. 254–264.

Voorhees, E.M. and Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In: *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pp. 316–323. ACM, New York, NY, USA.

Walker, D., Xie, H., Yan, K.-K. and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 6, p. P06010.

Walters, W.H. (2014). Do article influence scores overestimate the citation impact of social science journals in subfields that are related to higher-impact natural science disciplines? *Journal of Informetrics*, vol. 8, no. 2, pp. 421–430.

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, vol. 6, no. 4, pp. 700–711.

Waltman, L. (2015). NIH's new citation metric: A step forward in quantifying scientific impact? `https://www.cwts.nl/blog?article=n-q2u294`. [Online; accessed 10-Sep-2018].

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, vol. 10, no. 2, pp. 365–391.

Waltman, L. and van Eck, N.J. (2013). A systematic empirical comparison of different approaches for normalizing citation impact indicators. *Journal of Informetrics*, vol. 7, no. 4, pp. 833–849.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. and van Raan, A.F. (2011*a*). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, vol. 5, no. 1, pp. 37–47.

Waltman, L., van Eck, N.J., van Leeuwen, T.N., Visser, M.S. and van Raan, A.F.J. (2011*b*). Towards a new crown indicator: an empirical analysis. *Scientometrics*, vol. 87, no. 3, pp. 467–481.

Waltman, L., Yan, E. and van Eck, N.J. (2011*c*). A recursive field-normalized bibliometric performance indicator: an application to the field of library and information science. *Scientometrics*, vol. 89, no. 1, p. 301.

Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, vol. 94, no. 3, pp. 851–872.

West, J., Bergstrom, T. and Bergstrom, C.T. (2010). Big macs and eigenfactor scores: Don't let correlation coefficients fool you. *Journal of the American Society for Information Science and Technology*, vol. 61, no. 9, pp. 1800–1807.

West, J.D., Jensen, M.C., Dandrea, R.J., Gordon, G.J. and 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*, vol. 64, no. 4, pp. 787–801.

Wikipedia (2014). Lists of important publications in science. http://en.wikipedia.org/wiki/Lists_of_important_publications_in_science. [Online; accessed 19-Jan-2016].

Wuchty, S., Jones, B.F. and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, vol. 316, no. 5827, pp. 1036–1039.

Zheng, J. and Liu, N. (2015). Mapping of important international academic awards. *Scientometrics*, vol. 104, no. 3, pp. 763–791.

Zitt, M., Ramanana-Rahary, S. and Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, vol. 63, no. 2, pp. 373–401.

Zitt, M. and Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, vol. 59, no. 11, pp. 1856–1860.

Appendix

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

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 $\boldsymbol{x_0} = \boldsymbol{1/n}$ 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 $a_{ij} = (D(i))/(\text{od}(i))$. 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. $\boldsymbol{x_0} = \boldsymbol{P(i)}/|\boldsymbol{P}|$). 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:

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

where P(k) is the precision at cut-off k in the result set (described below) and rel(k) is a function that returns 1 if the element with rank k is relevant and 0 otherwise. 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

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

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.

Journal of Informetrics 12 (2018) 631–655

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

How to evaluate rankings of academic entities using test data

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[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 17 April 2018
Received in revised form 1 June 2018
Accepted 7 June 2018

Keywords:
Evaluating rankings
Test data
Cranfield paradigm
Significance testing

ABSTRACT

In the field of scientometrics, impact indicators and ranking algorithms are frequently evaluated using unlabelled test data comprising relevant entities (e.g., papers, authors, or institutions) that are considered important. The rationale is that the higher some algorithm ranks these entities, the better its performance. To compute a performance score for an algorithm, an evaluation measure is required to translate the rank distribution of the relevant entities into a single-value performance score. Until recently, it was simply assumed that taking the average rank (of the relevant entities) is an appropriate evaluation measure when comparing ranking algorithms or fine-tuning algorithm parameters.

With this paper we propose a framework for evaluating the evaluation measures themselves. Using this framework the following questions can now be answered: (1) which evaluation measure should be chosen for an experiment, and (2) given an evaluation measure and corresponding performance scores for the algorithms under investigation, how significant are the observed performance differences?

Using two publication databases and four test data sets we demonstrate the functionality of the framework and analyse the stability and discriminative power of the most common information retrieval evaluation measures. We find that there is no clear winner and that the performance of the evaluation measures is highly dependent on the underlying data. Our results show that the average rank is indeed an adequate and stable measure. However, we also show that relatively large performance differences are required to confidently determine if one ranking algorithm is significantly superior to another. Lastly, we list alternative measures that also yield stable results and highlight measures that should not be used in this context.

1. Introduction

New metrics and indicators for scoring academic entities are frequently proposed. To evaluate indicators on their utility for some task different approaches are taken. A metrics's mathematical soundness can be validated using axiomatic approaches (Altman & Tennenholtz, 2010; Bouyssou & Marchant, 2016). Two or more indicators can be compared to each other using correlation analyses. While this can yield some insight into proposed indicators, correlation analyses are problematic on their own (Thelwall, 2016) and can only be used as a comparison to some baseline (such as citation counts used as proxy for quality).

[*] Corresponding author at: Media Lab, Stellenbosch University, 7602 Matieland, South Africa.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

https://doi.org/10.1016/j.joi.2018.06.002
1751-1577/© 2018 Elsevier Ltd. All rights reserved.

Another approach is to use test data to evaluate ranking algorithms. One drawback of using test data is that its collection is expensive and time consuming. To decrease this effort lists of readily available data are often used as proxies for human judgements. Examples of such lists are: researchers that have received fellowship status at learned societies in recognition of their work (Dunaiski, Geldenhuys, & Visser, 2018; Nykl, Campr, & Ježek, 2015; Nykl, Ježek, Fiala, & Dostal, 2014); researchers that have won life-time contribution or innovation awards (Dunaiski, Visser, & Geldenhuys, 2016; Fiala, 2012; Fiala, Rousselot, & Ježek, 2008; Fiala & Tutoky, 2017; Gao, Wang, Li, Zhang, & Zeng, 2016; Nykl et al., 2014); and researchers that are frequently board members of prestigious journals (Fiala, Šubelj, Žitnik, & Bajec, 2015). For paper-level rankings, best paper awards or high-impact paper awards have been used (Dunaiski et al., 2016; Dunaiski & Visser, 2012; Mariani, Medo, & Zhang, 2016; Sidiropoulos & Manolopoulos, 2005).

We use the terms *metric* and *ranking algorithm* synonymously since they assign scores to academic entities that can be converted into a ranking (sorted list of entities with ascending ranks). When using test data (a subset of all entities considered important) to evaluate a ranking, some *evaluation measure* is needed to translate the rank distribution of the relevant entities into a single-value performance score. This paper deals with the evaluation measures and how they should be applied when evaluating ranking algorithms using test data.

Frequently, conclusions are based on simply using the average rank of the relevant entities as a performance score. This evaluation measure has been used to compare ranking algorithms to each other but also to draw conclusions about properties of the internal workings of the algorithms. For example, it has been used to judge whether self-citations should be included when computing impact scores of authors (Dunaiski et al., 2018; Nykl et al., 2014). Using the average rank as evaluation measure makes the assumption that if algorithm A ranks the important entities on average higher than algorithm B, then A must be better than B. However, it remains unknown whether the observed performance difference was obtained by algorithm A's superior ranking capabilities or was caused by outliers on a skewed rank distributions or simply occurred by chance. Moreover, how significant are the performance differences between the algorithms under investigation? Recently, alternative evaluation measures are adopted (Fiala & Tutoky, 2017) but the same problems remain: how confident are we about the obtained results?

In this paper we answer the above questions by addressing the following problem. The number of entities in a test data set is orders of magnitudes smaller than the number of authors or papers in real-world publication databases. Therefore the rank distribution of the test data entities is sparse and does not necessarily contain many high ranks. This situation causes many standard evaluation measures to become less effective. We show this by using methodologies from query-based information retrieval frameworks and adapting them for rankings of academic entities (Buckley & Voorhees, 2000; Sakai, 2006; Voorhees & Buckley, 2002).

Using our proposed framework, we analyse the discriminative power and stability of the evaluation measures on sparse rankings. The discriminative power is defined in terms of how well an evaluation measure distinguishes between significant and insignificant differences in rankings. The stability of a measure is based on its consistency of producing the correct results under changing conditions. In other words, we analyse an evaluation measure's general performance when the underlying data is changed and its volatility to rank biases.

The diagram in Fig. 1 depicts the workflow followed in this paper. Given a database of academic entities (papers or authors), they are ranked by metrics M_1 through M_k that assign scores to the entities. In Section 2 we discuss how these scores are converted into fair ranks. The next step is to extract the ranks of relevant entities of a test data set, in this case 'Test Set 1'. We describe the different test data sets used in this paper in Section 3 and outline the motivation behind this paper. Section 4 describes the most common evaluation measures in the context of academic entities and how they can be adjusted for percentile rankings. Based on the rank distribution of the relevant entities, the evaluation measures are used to compute performance scores for the metrics. We then formulate the framework of how these evaluation measures are evaluated (Section 5). In Section 6 we discuss the results of this second-order evaluation.

We make the following contributions:

- We propose a framework for evaluating evaluation measures based on rankings of academic entities that are part of test data. Using this framework the stability and discriminative power of the most common evaluation measures are analysed.
- The proposed methodology provides the capability of computing significance levels associated with performance differences between ranking algorithms.
- We show that simple measures such as the average or median rank have high discriminative power and are stable evaluation measures.
- We find that using permille rankings does not improve the performance of evaluation measures in general except for the nDCG measure which should only be used with permille rankings.
- Our results show that a "one size fits all" evaluation measure does not exist and that appropriate measures have to be chosen carefully based on the underlying data.

2. Converting scores to ranks

Ranking algorithms and impact indicators usually produce scores that are associated with entities. However, scores from different metrics are not directly comparable and have to be converted to ranks first. An entity with a larger score usually indicates that it is "better" than entities with smaller scores produced by the same metric. Therefore, the output of metrics

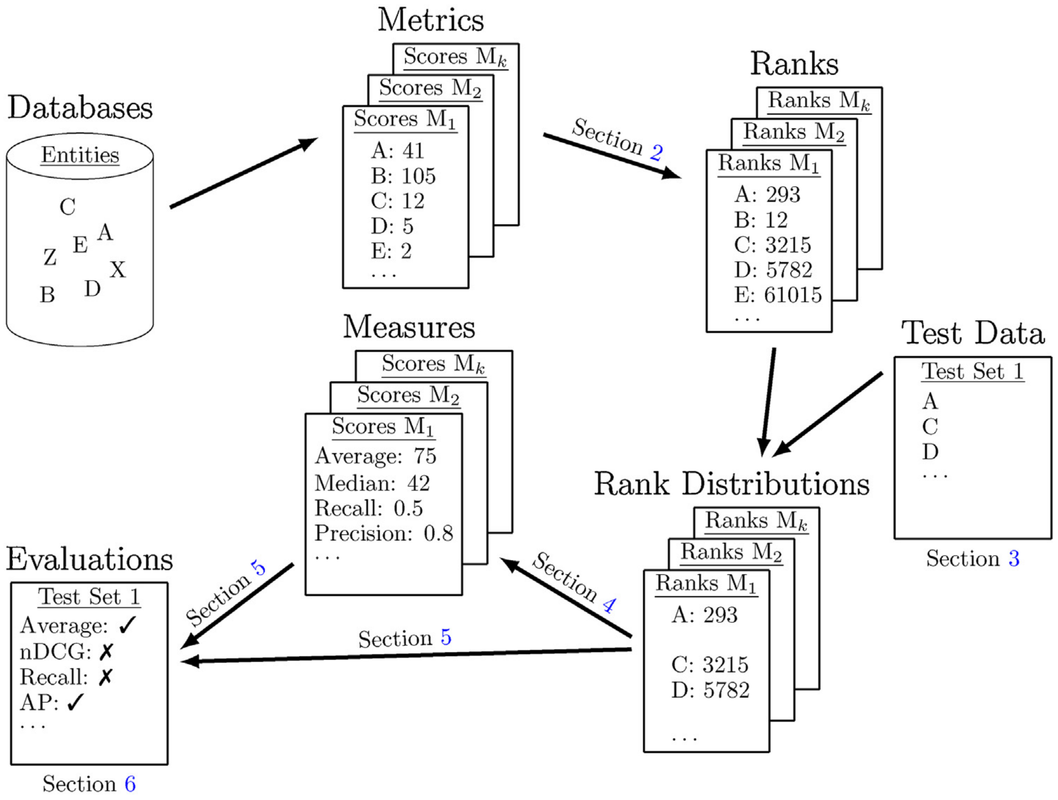


Fig. 1. Overview of the process followed in this paper. Entities from real-world databases comprising papers and authors are ranked by metrics (M_1–M_k) that assign scores which have to be translated into fair ranks. The ranks of the relevant entities that belong to the test data are extracted and produce a rank distribution for each ranking metric. These are used to compute performance scores for each metric. Based on these rank distributions the evaluation measures are evaluated using the framework proposed in this paper.

can be transformed into ordered lists of ranks in ascending order where a higher rank, starting at 1, is “better” than lower ranks. When referring to rank values, we use the adjectives *small* and *large* to describe the value where a smaller rank value (higher rank) is “better” than a larger rank value (lower rank).

For fair comparisons across all metrics under consideration, it is necessary that only entities are used that have scores assigned by all metrics. This is especially important if a subset of entities is used as test data to evaluate the metrics’ performances. For example, if a metric assigns scores to many more entities than others it is easy to imagine that the average rank of the subset of entities is much lower (larger value) compared to the metrics that score fewer entities.

A set of scores is not guaranteed to be well-ordered since entities may have equal scores, which happens often with discrete-value metrics such as the *h*-index (Hirsch, 2005). Below we describe different approaches of handling ties when converting scores into a list of ranks. Assume that some metric assigns scores $\{25, 24, 24, 20\}$ to a set of entities with two entities having equal scores of 24. The standard approach assigns the same ranking value to entities with equal scores, after which a gap is left. The size of the gap is one less than the number of preceding ties. Using the above example, the ranks of the entities would be $\{1, 2, 2, 4\}$. Therefore, an entity’s rank is 1 plus the number of entities that are ranked above it. Alternatively, with the *ordinal ranking* approach, entities with equal scores are assigned distinct rank values that are randomly drawn from the set of ranks that they share. For the above example, the ranks of the entities could either be $\{1, 2, 3, 4\}$ or $\{1, 3, 2, 4\}$. Lastly, with *fractional ranking*, entities that compare equal are assigned the same rank value which is the mean rank of the ranks that they would have under the ordinal ranking approach. Therefore, the entities’ ranks using this approach would be $\{1, 2.5, 2.5, 4\}$.

For ordinal ranking, the random assignment of ranks for entities with equal scores can lead to unfair rank assignments depending on whether relevant entities are chosen before non-relevant entities or vice versa. This is shown by the example in Table 1 . The scores produced by some metric are given at the top of the table ranging from 25 to 12. Relevant entities are highlighted in grey. The last three columns show the average rank, the median rank, and the average precision of the relevant entities’ ranks.

Row one shows the ranks of the entities when the standard ranking approach is used to convert scores to ranks. It is easy to see that using this approach leads to unfair comparisons between metrics when, for example, the average or median rank is used for evaluation. The more ties a metric produces, the better it will perform on average.

Rows two and three show two different rankings obtained when the ordinal ranking approach is used. Even though the entities’ scores are identical, the average and median rank of the relevant entities produced by ‘Ordinal 1’ are 4.0 and 3, which differ from 5.0 and 5 produced by ‘Ordinal 2’.

Table 1
Different ranking approaches when converting scores to ranks. The scores are given in the first row where relevant entities are highlighted in grey. The columns 'Avg.', 'Med.', and 'AP' give the average rank, the median rank, and the average precision computed from the ranks of the five relevant entities.

Score	25	24	24	24	24	20	20	12	Avg.	Med.	AP
Standard	1	2	2	2	2	6	6	8	3.8	2	0.958
Ordinal 1	1	5	2	3	4	6	7	8	4.0	3	0.857
Ordinal 2	1	2	4	5	3	7	6	8	5.0	5	0.659
Fractional	1	3.5	3.5	3.5	3.5	6.5	6.5	8	4.5	3.5	0.734

One solution to this problem is to convert the set of scores multiple times, each time randomly assigning ranks to the entities with equal scores. The evaluation measure (e.g., computing the average rank of relevant entities) can be computed for each list and averaged. This will lead to a fairer evaluation of the metric but it does not solve the problem of obtaining fair and balanced rankings if they were required. Furthermore, it raises the question of how many times a new random list of ranks has to be sampled before a fair evaluation score is obtained.

The last row shows the ranks of the entities if fractional ranking is used. It is fairer than standard ranking when evaluation measures such as the average or median rank is used. Moreover, it overcomes the aforementioned problem of ordinal rankings. However, as shown in Section 4.2, caution is warranted when using the standard or fractional ranking approaches with other evaluation measures since they can lead to inconsistent results.

Fiala and Tutoky (2017) convert scores into permille ranks where entities are assigned a rank between 1 and 1000 based on which permille group they fall into compared to all entities. In other words, an entity is assigned the rank *k* if its score falls into the top-*k*‰ of all scores. Since scores produced by metrics are not uniformly distributed, permille ranks have to be based on ranks and not on scores. One of the other approaches discussed in this section has to be used before converting ranks into permille ranks. Additional care has to be taken when assigning permille ranks to entities when equal ranks overlap permille boundaries. We use the approach described by (Waltman & Schreiber, 2013) to average ranks over boundaries which leads to fractional ranks.

3. Evaluating academic ranked retrieval results

We use two publication databases. The first is a copy of the Digital Library (ACM, Inc., 2014) of the Association for Computing Machinery (ACM). It contains 1.8 million papers published up to March 2015 in periodicals and proceedings from the field of computer science. The second database is the Microsoft Academic Graph (MAG) (Microsoft, 2017) from 2016 which is cross-disciplinary and comprises over 126 million papers.

We use four different test data sets that consist of different types of relevant entities and have been matched to both the ACM and MAG databases:

- **ACM fellows**: 930 authors that have received fellowship accreditation by the ACM have been matched to both the ACM and MAG databases.
- **LCA authors**: 393 and 507 authors from 27 different award committees that have received long-term contribution awards (LCA) have been matched to the ACM and MAG databases, respectively.
- **BPA papers**: 568 (ACM) and 587 (MAG) papers that have received best paper awards (BPA) collected from 36 different conferences or learned societies.
- **HI papers**: 443 (ACM) and 406 (MAG) papers from 30 different venues that have won high-impact (HI) awards.

Given test data comprising entities we assume to be relevant, we want to evaluate a metric's performance based on the ranks it assigns to them. Let *Rel* be a set of relevant entities. A perfect metric would rank the *Rel* entities in the top |*Rel*| ranks. However, the difficulty with a ranked list of real-world data, which is usually orders of magnitudes larger than the set of test data, is that the relevant entities are spread out substantially and not necessarily very high in the rankings.

For example, Fig. 2 shows the rank distribution (up to rank 5000) of the ACM fellows and LCA authors when citation counts are used as the ranking metric on the ACM database. The distribution is long-tailed with 55% of relevant entities ranking lower than rank 1000. On average the ACM fellows are more prolific than the average author in the ACM database. Furthermore, papers that are included in the ACM database generally belong to the same fields in which the ACM fellows produced work and received recognition for. Nevertheless, some ACM fellows rank relatively low. It is easy to imagine that this problem is exacerbated when the rank distribution is based on a multi-disciplinary database where the ACM fellows compete against prolific researchers from other disciplines. The same argument is valid for the LCA authors.

We are also interested in whether some metrics identify ACM fellows that are less prolific and rank them higher compared to equally prolific researchers that have not received recognition for their work. In other words, how good is a metric in discriminating between relevant entities and non-relevant entities at lower ranks? For example, do the ranks of the ACM fellows with lower citation counts improve in general when the *h*-index is used for comparisons instead? Therefore we conjecture that it is important for evaluation measures to use a substantial part of the rank distribution in order to discriminate between different ranking algorithms.

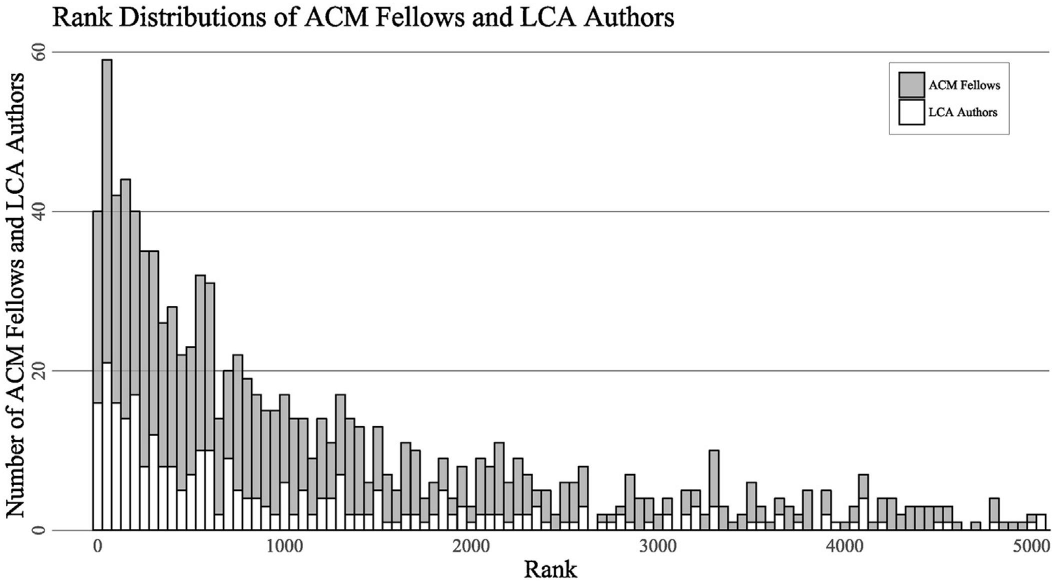


Fig. 2. Rank distribution up to rank 5000 of the ACM fellows and LCA authors according to citations counts on the ACM database.

The same situation applies to the paper level where some BPA papers might not have received a lot of citations and therefore are ranked relatively low. Yet we are still interested in whether certain ranking algorithms can identify those papers regardless and rank them higher than papers that have equal citation counts but have not won a best paper award.

4. Evaluation measures

The most elementary way to measure ranked lists of entities is to compute the **average** or **median** rank of all relevant entities. Since it makes sense intuitively that if one metric ranks the relevant entities higher on average than another metric, the former metric should be regarded as better than the latter. However, the average rank can easily be dominated by a small number of outliers in a skewed distribution. Furthermore, it is easy to attribute too much significance to a small change in the mean which ultimately might not be proportional to the difference in rankings and could be an artefact of random noise. To a lesser degree the same concerns pertain to the median as an evaluation measure. However, the median rank measure has less discriminative power than the average rank measure when two rankings are similar.

In this paper we also analyse two rudimentary measures that only consider the rank of a single relevant entity. The **Min** and **Max** measures simply use the smallest and largest rank value of the relevant entities as performance scores. Therefore, the smaller the score produced by these measures the better the ranking is regarded. However, both measures are susceptible to outliers since only a single outlier is required to bias the score.

The two most frequently used measures for evaluating retrieval results are precision and recall. Most other measures discussed in this section are based on these two measures. **Precision** (P) is defined as the fraction of retrieved entities that are relevant, and **recall** (R) is the fraction of relevant entities that are retrieved. The problem with both these measures is that it is assumed that some relevant entities might not be retrieved by an algorithm. We therefore need a cut-off threshold *n* (e.g., the top 100 ranked entities) and count the number of relevant entities that are ranked up to this cut-off.

Let *tp* (true positive) and *fp* (false positive) respectively be the number of relevant and non-relevant entities ranked in the top *n* ranks. Furthermore, let *fn* (false negative) be the number of relevant entities that are not ranked in the top *n* ranks. For completeness, let *tn* (true negative) be the number of non-relevant entities that are not ranked in the top *n* ranks, which is not very useful in the context of ranked lists of academic entities since it will usually be a very large number. Using these terms, precision and recall are defined as:

$$\mathrm{P} = \frac{tp}{tp+fp} \qquad \mathrm{R} = \frac{tp}{tp+fn} \tag{1}$$

If *Rel* is the set of relevant entities, then precision and recall at cut-off threshold *n* are defined as follows:

$$\mathrm{P@}n = \frac{|\{rel \in Rel \mid rank(rel) \leq n\}|}{n} \qquad \mathrm{R@}n = \frac{|\{rel \in Rel \mid rank(rel) \leq n\}|}{|Rel|} \tag{2}$$

where *rank*(*rel*) is the rank of the relevant entity *rel*.

Considering the rank distribution in Fig. 2 again and setting *n* = 100, the majority of relevant entities would be ignored. This is not reasonable since we are still interested in where the lower ranked entities rank in comparison to other metrics. Setting *n* to a very large cut-off threshold also does not work since the precision would tend towards zero given the large number of non-relevant entities. In contrast, recall is a non-decreasing function and can always be increased by simply

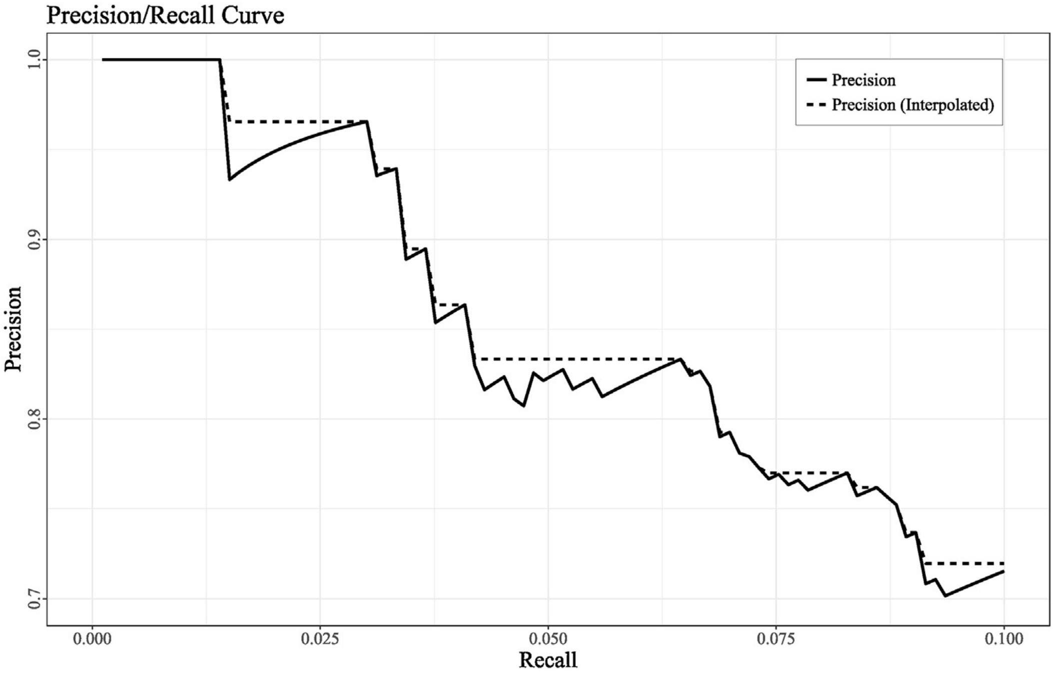


Fig. 3. A typical precision-recall curve using ACM fellows as relevant entities. Citation counts are used as ranking metric on the ACM database. The precision and recall values are plotted only up to rank 130, where a recall level of 0.1 is achieved (93 of the 930 ACM fellows have ranks higher than 130). The solid curve indicates the precision and recall values at various cut-off thresholds while the dashed curve shows the corresponding interpolated precision.

choosing larger cut-off thresholds. Furthermore, it is problematic to define an arbitrary cut-off since it influences the results between different test sets and databases.

When evaluating ranked retrieval results it is possible to compute precision and recall of the top n entities at various cut-offs. For each set the precision and recall values can be plotted yielding a **precision-recall curve** as indicated by the solid curve in Fig. 3. In this example, citation counts are used as the impact metric to rank the ACM fellows on the ACM database. The curve has the typical saw-tooth shape since when the $(k+1)$-th ranked entity is not relevant, the recall value stays the same as for the top k entities, but the precision will decrease. However, if this entity is relevant, then both precision and recall increase.

A common method to smoothen out this curve is to compute an **interpolated precision**. Let $\mathrm{P}_{inter}@r$ be the interpolated precision at some recall level r which is the highest precision found for any other recall level $r' \geq r$. More formally:

$$\mathrm{P}_{inter}@r = \max_{r' \geq r} \mathrm{P}@r' \tag{3}$$

The interpolated precision is given by the dashed curve in Fig. 3. In general information retrieval contexts, interpolating precision is justified by arguing that users are typically prepared to look at a few more results in order to increase the percentage of viewed results that are relevant. To convert the precision-recall curve into a single-value performance score, one can compute the average (interpolated) precision at predefined recall levels (usually at levels [0.0, 0.1, . . . 1.0] which is referred to as the 11-point interpolated average precision (Manning, Raghavan, & Schütze, 2008)). Alternatively, the area under the precision-recall curve can be computed (PR).

The **average precision** (AP) is a value obtained by computing the average of the non-interpolated precision scores at each rank where a relevant entity is retrieved and therefore factors in precision at all recall levels. However, for many applications such as web search results only the top results are important. For these types of applications, it is only important how well elements are ranked within the top 10 or 20 results. This is done by choosing cut-off thresholds at small numbers of retrieved results.

This is incorporated in a variant of AP called **average precision at** n (AP@n) with a cut-off threshold n. AP@n is the average precision at each rank up to n and is defined as follows:

$$\mathrm{AP}@n = \frac{1}{\min(|Rel|, n)} \cdot \sum_{k=1}^{n} \mathrm{P}@k \cdot \mathrm{isrel}(k) \tag{4}$$

where isrel(k) is a binary function that returns 1 if the entity with rank k is relevant and 0 otherwise. Therefore, AP@n measures how many relevant entities are returned by a query in the top n ranked entities and their average precision.

For example, consider three ACM fellows that, according to their citation counts, are ranked in positions 1, 5, and 11 in a list of otherwise non-fellows. The AP@10 for citation counts using the three ACM fellows as the test set would then be

(1/1 + 2/5)/3 = 0.56. AP is biased towards the top of the rankings since placing the first relevant entity in the first rank weighs twice as much as placing it in the second rank.

When two or more lists of rankings are available for evaluating a ranking algorithm, an evaluation score can be computed using each list separately and averaged to produce a final performance score. In the information retrieval field, rankings are associated with *queries* that are processed by some recommendation algorithm. For continuity we retain this terminology.

The **mean average precision at** *n* (MAP@*n*) is the mean AP@*n* of a set of *N* queries:

$$\text{MAP@}n = \frac{1}{N} \cdot \sum_{i=1}^{N} \text{AP@}n(i) \tag{5}$$

where AP@*n*(*i*) is the average precision at *n* for query *i*. The AP value is an approximation of the area under the non-interpolated precision-recall curve. Therefore, MAP is the average area under the precision-recall curves for a set of *N* queries.

In text-based information retrieval environments, the calculated MAP scores normally vary widely across queries when evaluating a single ranking algorithm (Manning et al., 2008, p. 161). In general it has been observed that there is more agreement in MAP for a specific query across different ranking algorithms than between MAP scores for different queries using the same algorithm. Therefore, to evaluate an algorithm rigorously across varying queries, the test set must be large and diverse enough to be representative of the algorithm's effectiveness across different queries. In the case of ranking academic entities this concern is somewhat alleviated since all relevant entities are treated the same and their retrieval does not depend on the formulation of the query. For example, if the set of ACM fellows is divided into groups to represent different queries, the behaviour of the ranking algorithms does not change depending on the ACM fellows that are contained in a group.

The advantage of using cut-off thresholds is that the size of the test set does not need to be known. The disadvantage however is that it does not average well (Buckley & Voorhees, 2000), since the total number of relevant entities for a query has a strong influence on P@*n*. This concern is alleviated if the same test set is used for identical queries evaluating different ranking algorithms.

An alternative, which overcomes the problem of averaging over different queries, is ***R*-precision**. It requires having a set of known relevant entities *Rel*, from which we calculate the precision of the top-|*Rel*| entities returned. *R*-precision adjusts for the size of the set of relevant entities: A perfect algorithm could score 1 on this measure for each query, whereas, even a perfect algorithm would only achieve 0.3 when *P*@20 is used and there were only 6 entities relevant to a query. Therefore, this measure can safely be averaged across different queries. Given that *r* relevant entities are found in the top-|*Rel*| returned entities, *R*-precision describes one point on the precision-recall curve, which is the break-even point where precision and recall are both *r*/|*Rel*|.

An alternative to using the precision-recall curve for evaluations is the "Receiver Operating Characteristics" (ROC) curve. The **ROC curve** plots the true positive rate (tpr_i) against the false positive rate (fpr_i) at various cut-off intervals *i* defined in the set *I*. The true positive rate is the same as recall (*tp*/(*tp* + *fn*)) and the false positive rate is given by *fp*/(*fp* + *tn*). For the cut-off interval *i*, the tpr_i and fpr_i can be defined as:

$$tpr_i = \frac{|\{rel \in Rel \mid rank(rel) \le i\}|}{|Rel|}; \qquad fpr_i = \frac{i - |\{rel \in Rel \mid rank(rel) \le i\}|}{\max(I)} \tag{6}$$

Using the ACM fellows on the ACM database again, Fig. 4 shows the ROC curves for three author ranking metrics. A ROC curve typically starts at the bottom left of the graph and ends at the top right side. The ROC curve for a well performing algorithm climbs quickly on the left-hand side and plateaus near 1.0. It is helpful to plot a line with a slope of 1 for comparison which corresponds to random assignments of ranks. Typically, the area under the ROC curve (ROC AUC) is computed to obtain a single-value measure and can be seen as the analog of the MAP value for the precision-recall curve. As can be seen in Fig. 4, both the *tpr* and the *fpr* do not reach 1.0 if one or more relevant entities rank higher than max(*I*).

Instead of defining various cut-off intervals for the ROC computations, only a single cut-off threshold *n* can be used (ROC@*n*). Then at every level ($rank(rel_i)$) of the *i*-th relevant entity a capped true positive rate and false positive rate is computed as follows:

$$\begin{aligned} tpr_i &= \begin{cases} tpr_i & i \le n \\ tpr_n & i > n \end{cases} \\ fpr_i &= \begin{cases} 0 & rank(rel_i) = i \\ \dfrac{rank(rel_i) - i}{(rank(rel_i) - i) + n - i} & i \le n \\ fpr_n & i > n \end{cases} \end{aligned} \tag{7}$$

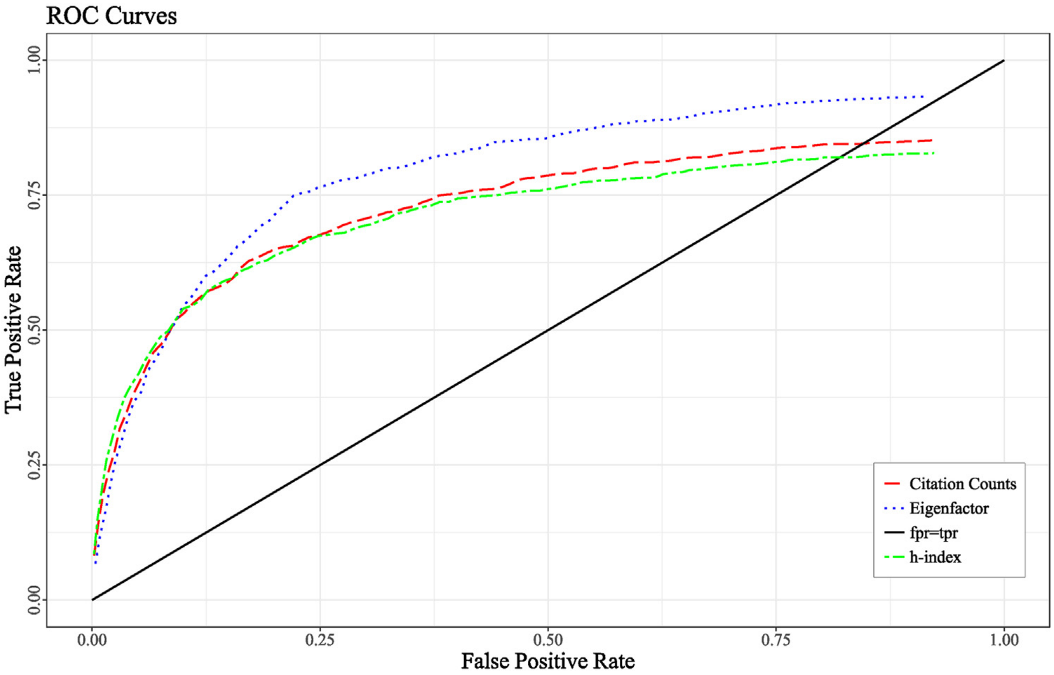


Fig. 4. ROC curves of using citation counts, author-level Eigenfactor, and the *h*-index as metrics to rank the ACM fellows on the ACM database. The area under the curves are respectively 0.662, 0.714, and 0.652.

The **discounted cumulative gain** (DCG) measure (Järvelin & Kekäläinen, 2002) is a more recent metric which can also be used when test data is non-binary (i.e., relevance scores are assigned to the relevant entities). If *rel*(*e*) is the relevance score associated with entity *e*, then DCG with cut-off threshold *n* is defined as:

$$\mathrm{DCG@}n = \sum_{e \in Rel|rank(e) \leq n} \frac{rel(e)}{\log_2(1 + rank(e))} \tag{8}$$

In order to obtain comparable values, DCG@*n* has to be normalised.[1] This is done by dividing the obtained DCG@*n* score by the DCG@*n* for an optimal ranking where the relevant documents are ordered in descending order of relevance and ranked in the top rankings. This is the **normalised discounted cumulative gain** (nDCG) measure.

Similar to the average precision, the nDCG measure biases higher ranks and penalises lower ranks by a logarithmic factor. AP can be summed over all ranks while the nDCG measure should be used with a cut-off threshold because the nDCG distribution is long-tailed and relevant entities at lower ranks would outweigh the utility of relevant entities at higher ranks.

4.1. *Evaluation measures for percentile rankings*

The average and median rank of relevant entities can still be used on percentile rankings directly. However, measures based on precision or recall, as well as the nDCG measure require adjustments when used on percentile or permille rankings.

In general, ranks can be assigned to any predefined number of buckets *b*. Here we describe permille rankings where the number of buckets is set to 1000. Fig. 5 depicts the ACM fellows ranked by citation counts and converted into permille ranks ($b = 1000$). $T_i = i \cdot (T/b)$ gives the total number of entities up to and including bucket *i*, where *T* is the total number of ranked entities. R_i indicates the number of relevant entities that have ranks higher than or equal to *i*. For example, of the 625 authors with a rank of 1, 294 are ACM fellows. Similarly, 465 ACM fellows have ranks of 2 or higher.

We suggest that for evaluation measures without cut-offs such as AP, the precision values are calculated at every bucket boundary. The precision at bucket boundary *i* is simply $\mathrm{P@}(T_i) = R_i/T_i$. When using a cut-off threshold *n* that does not fall onto a bucket boundary, only a proportion of a bucket should be considered. For example in Fig. 5, the cut-off $n = 2000$ falls into the 4*th* bucket.

For the following definitions let *i* be the largest value such that $T_i \leq n$. Let *ratio*(*n*) be the proportion of the *i-th* bucket that has to be considered when using the cut-off *n*:

$$ratio(n) = \frac{n - T_i}{T_{i+1} - T_i} \tag{9}$$

[1] Sometimes the numerator for DCG is defined as $2^{rel(e)} - 1$ instead of *rel*(*e*). For non-binary graded relevance scores this version places greater emphasis on relevant entities but does not change the results when the relevance values of the entities are binary, i.e., $rel(e) \in \{0, 1\}$.

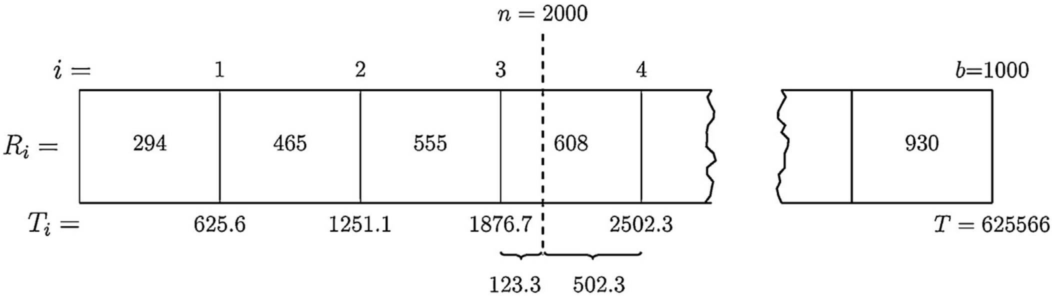


Fig. 5. Illustrative example of permille rankings. The ACM database comprises a total of T authors that are partitioned into $b = 1000$ buckets. T_i indicates the number of authors up to and including bucket i. R_i indicates the number of relevant entities (ACM fellows) that have rank values smaller than or equal to i.

The precision at cut-off n is then defined as

$$\text{P@}(n) = \frac{R_i + ratio(i) \cdot (R_{i+1} - R_i)}{n} \tag{10}$$

and the average precision with cut-off n is adjusted as follows

$$\text{AP@}(n) = \frac{\sum_{k=1}^{k<i} \text{P@}(T_k) + \text{P@}(n)}{\sum_{k=1}^{k<i}(1) + ratio(n)} \tag{11}$$

where the denominator is the number of buckets that are fully used plus the ratio of the partially used bucket, i.e., 3 + (123.3/625.6) = 3.197. The definitions for the other measures, such as Recall@n, nDCG@n, or ROC@n are derived similarly.

4.2. The problem with precision and rankings with ties

As mentioned in Section 2 when converting scores to ranks, the fractional ranking approach assigns the average rank to each entity that scored equally. This can result in a situation where the average rank value is smaller than the total number of ranks considered up to this point. For example, assume that the three top scores are {25, 24, 24} of entities that are all relevant. Their corresponding fractional rank values are {1, 2.5, 2.5}, where the rank value (2.5) in the third position is smaller than its rank position (3). This leads to precision values above 1 which is infeasible. In this example, the precision values at the first three cut-off thresholds {P@1, P@2, P@3} are therefore {1, 0.8, 1.2}. The same is true for the standard ranking approach where the ranks would be {1, 2, 2} which results in the same three precision values as with the fractional ranking approach.

In Fig. 6 the first part of the precision-recall curve is plotted where the *h*-index is used as the metric to rank the ACM fellows on the ACM database. One can see that the precision at certain ranks (red dotted curve) spikes above 1. One solution is to cap the precision at 1 indicated by the green dashed curve and interpolate these values according to Eq. (3) which is indicated by the blue solid curve in Fig. 6.

Let $\text{P}_{adj}@r = \min\left\{\text{P}@r, 1\right\}$ be the adjusted precision at some recall level r. Then the adjusted and interpolated precision at r is

$$\text{P}_{adj+inter}@r = \max_{r' \geq r} \text{P}_{adj}@r' \tag{12}$$

5. Experimental framework of evaluating evaluation measures

In a typical information retrieval environment, the task of an information retrieval system is to return the most relevant documents pertaining to a user's information need. This is achieved by ranking more relevant documents above less or non-relevant documents. The system typically indexes a collection of documents that are grouped into a number of topics. When executing a user's query, the system chooses documents from these topics and returns them as a ranked list according to their estimated relevance.

To judge the effectiveness of an information retrieval system, ground truth data is required. In information retrieval settings this forms part of a *test collection*. It consists of a corpus of documents (D), a set of topics (T), and relevance judgments (R) for documents associated with a controlled number of queries for all topics. A system's effectiveness is evaluated based on where it ranks the relevant documents associated with queries. Therefore, evaluation measures are required that measure effectiveness based on rankings of relevant documents.

For a test collection (D, T, R) a system performs its function and retrieves a document list (L_t) for each topic $t \in T$. An appropriate evaluation measure ($eval(R, L_t)$) is used to judge the quality of the retrieved document lists separately, producing an evaluation score for each topic. These are combined (usually) using the arithmetic mean to produce a single performance value. The evaluation measure's purpose is to decide how well a system performs its function, to determine the best function of a system, or to compare different systems to each other.

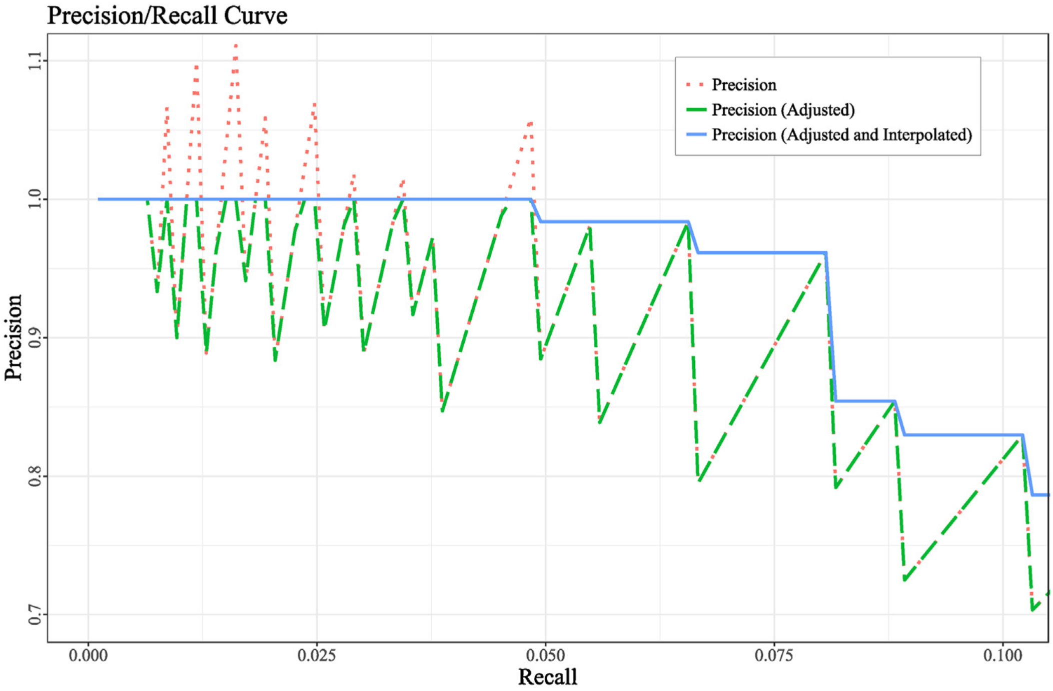


Fig. 6. Illustration of how the precision at certain recall levels (dotted curve) can exceed 1.0 when entities with equal scores are assigned ranks using the standard or fractional ranking approaches. The data used in this plot are the ranks of ACM fellows when using the *h*-index as ranking metric on the ACM database. The adjusted precision curve (dashed) caps the precision values at 1.0, while the solid curve shows the interpolated precision curve based on the adjusted precision values.

The Cranfield paradigm puts forth a set of rules for fair evaluation of retrieval systems (Cleverdon, 1967). It states that evaluation requires a fixed set of documents (corpus), a fixed set of topics (information needs) and that the relevance judgments are complete, i.e., that for every query all retrieved documents have assigned relevance judgements.

The Cranfield paradigm also makes three assumptions. The first assumption is that relevance is approximated by topical similarity. This implies that all relevant documents are equally relevant and independent from each other. Furthermore, it implies that the user information need is static. The second assumption is that a single set of judgements for a topic is representative of the user population. The last assumption is that the list of relevant documents for each topic is complete, that all relevant documents are known.

However, in general these assumptions are not true, which makes the evaluation of information retrieval systems a noisy process (Voorhees, 2002). To decrease the noise, the now standard experimental design sets forth that each retrieval system produces a ranked list of documents for each topic in the test collection. A retrieval system's effectiveness for a single topic is computed as a function of the ranks associated with the topic's relevant documents. The effectiveness of the retrieval system as a whole is computed as the average effectiveness over all topics in the test collection.

In the case of an academic corpus, natural topics exist that can be delineated by academic disciplines, language or narrow fields of research. However, these topics are not applicable in the context of ranking academic entities. Internally the algorithms may utilise contextual information such as computing the similarity between two papers or normalising a paper's score based on the field it belongs to. However, the list of ranked entities returned by a ranking algorithm is always the complete set of entities and not a subset which depends on the topic of the query.

Therefore, all queries are identical and are also independent of topics, i.e., a query does not take any context into consideration such as “which papers are the most similar to machine learning?” Yet we may still ask the question “what are the top papers in the field of machine learning?” In systems that make use of topics, the effectiveness of retrieval systems may vary widely across topics (Manning et al., 2008), since some systems might perform poorly when a topic is broadly defined and work well for precisely defined topics. In this case, the more topics are used, the more confident the experimenter can be in its conclusions.

It should be noted that evaluation measures that have been developed vary widely in stability. For example, measures based on very little data such as P@1 are very noisy compared to the average precision measure which is more stable (Voorhees, 2002). In general, requiring a larger difference between scores before considering the retrieval systems to be significantly different increases reliability at the cost of not being able to discriminate between as many systems.

A potential difficulty for evaluation measures applied on ranks of academic entities is that most ranking algorithms and indicators are ultimately based on citations. We therefore hypothesise that ranked lists, as produced by citation-based metrics, are inherently similar and therefore place a higher burden on evaluation measures' discriminative powers.

The evaluation design described above consists of three interrelated variables: the number of topics used, the evaluation measure used, and the difference in scores required to consider one information retrieval system better than another. In

Table 2
A depiction of a test collection consisting of 25 topics (t) with 10 different query expressions (i) resulting in unique queries $Q_{i,t}$. The experimental design proposed by Buckley and Voorhees (2000) uses query sets S_i consisting of one unique query per topic. In this design query expressions are created by different systems which is indicated by the query type.

Query Type	Topic 1	Topic 2	. . .	Topic 25	Query Set
1 {	$\mathcal{Q}_{1,1}$ = B. Obama	$\mathcal{Q}_{1,2}$ = Tree	. . .	$\mathcal{Q}_{1,25}$ = PC	$\mathcal{S}_1$
	$\mathcal{Q}_{2,1}$ = Barack	$\mathcal{Q}_{2,2}$ = Bonsai	. . .	$\mathcal{Q}_{2,25}$ = Laptop	$\mathcal{S}_2$
2 {	⋮	⋮	⋱	⋮	⋮
⋮					
5 {	$\mathcal{Q}_{10,1}$ = President	$\mathcal{Q}_{10,2}$ = Plant	. . .	$\mathcal{Q}_{10,25}$ = Computer	$\mathcal{S}_{10}$

Table 3
Illustrative example of counts of the number of times the system of the row was better than, equal to, or worse than the system of the column. A table is shown for each of the two evaluation measures Eval1 and Eval2.

Eval1	B			C			D		
A	10	0	0	8	1	1	2	3	5
B				5	1	4	4	3	3
C							5	2	3

Eval2	B			C			D		
A	8	1	1	6	2	2	2	5	3
B				4	2	4	3	4	3
C							4	3	3

Section 5.1 we describe the well-studied experimental frameworks for evaluating evaluation measures in the context of typical text-based information retrieval systems. In Section 5.2 we propose an adapted framework for evaluating evaluation measures based on systems that rank academic entities.

5.1. The traditional experimental framework

In typical text-based information retrieval settings, a *query* is an expression of a topic that is processed by the retrieval system to retrieve (ideally) relevant documents associated with the topic and the query. Using different queries for the same topic affects the retrieval behaviour since some queries are better expressions of a topic than others. By varying the expressions of a topic and observing how the evaluation measures' results vary, one can calculate an error associated with the measure. A *query set* is a collection of queries, one for each topic.

Table 2 depicts this graphically. The test collections consist of 25 topics and 10 queries for each topic. The query $Q_{i,t}$ is a unique expression of a query for topic t where the query set S_i contains one unique query for each topic.

5.1.1. Error and tie rate

Buckley and Voorhees (2000) first introduced a method to estimate the sensitivity of evaluation measures and a notion of computing their discriminative power. They define an *error rate* which intuitively can be interpreted as the percentage of time a measure comes to the wrong decision when evaluating two systems. The error rate therefore indicates a lack of stability. They also define a *tie rate* which is the percentage of time a measure fails to decide which of two systems performs better. The tie rate indicates a lack of discriminative power.

To compare two evaluation scores to each other they define a *fuzziness* value which is the percentage (absolute) difference between two scores such that if the difference is smaller they are considered equivalent. For their approach they pick a query set S_i and compute the mean of the evaluation measure over the query set for each retrieval system. For each pair of retrieval systems they record the number of times that one system is better, equal, or worse than the other system. For a query set, two systems are considered equal if the evaluation score only deviates from the mean over all systems within a percentage defined by the fuzziness value. This is repeated for all query sets which results in a decision table for an evaluation measure.

To illustrate such a table assume that four retrieval systems (A, B, C and D) and two evaluation measures (Eval1 and Eval2) are used. Table 3 contains some made-up decision values for Eval1 and Eval2. The three values in each table entry show, respectively, the counts of the number of times the evaluation measure scored the system of the row better than, equal to, or worse than the system of the column.

Buckley and Voorhees (2000) assume that for each pair of retrieval systems the correct answer is given by the greater of the better-than and worse-than values. Then it follows that the lesser of these two values is the number of times an evaluation result is misleading or in error. Therefore the error rate of an evaluation measure is the total number of errors across all system pairs (*X*, *Y*) divided by the total number of comparisons:

$$Error\ Rate = \frac{\sum \min(|X > Y|, |Y > X|)}{\sum(|X > Y| + |X < Y| + |X == Y|)} \tag{13}$$

Considering the example in Table 3 the error rate for Eval1 is 13/60 = 0.217 and for Eval2 is 15/60 = 0.250. The error rate can never be more than 50% and if the error rate exceeds 25% Buckley and Voorhees (2000) assume that random effects dominate the calculation.

The tie rate is the average percentage of the number of times a measure could not decide which of two systems was better given the fuzziness value. From the above decision table, the tie rate for a measure is the number of indecisions between system pairs divided by the total number of decisions. For the measures Eval1 and Eval2 the tie rates are respectively 10/60 = 0.167 and 17/60 = 0.283.

Using this approach, two values have to be considered when judging the effectiveness of evaluation measures. For example, a measure might have a very low error and tie rate, which indicates that the measure has high stability and good discriminative power. The two values can also diverge where a measure can have a low error rate but a high tie rate, indicating that the measure has low discriminative power (high tie rate) but that the few remaining decisions were decided mostly correctly.

To obtain average error and tie rates with deviation intervals, Buckley and Voorhees (2000) recompute the values multiple times by creating 50 different sets of permuted query sets. They evaluate P@*k* at various cut-off thresholds, Recall@1000, Precision@(50% recall), *R*-precision, and AP. They use the results of 9 different retrieval systems submitted to the Query Track TREC-8 conference[2] and find that P@1 has the highest average error rate (14.3%, $\sigma = 1.3$). They also find that AP has a lower error rate than average precision at various cut-off thresholds except when P@1000 is used. The measure with the lowest error rate is Recall@1000 with 0.6% ($\sigma = 0.2$) and a tie rate of 20.8%. Looking at the tie rates, they find that AP and Precision@(50% recall) have the highest discriminative power with tie rates of 12.8% and 11.4%, respectively. Additionally, Buckley and Voorhees (2000) analyse how varying the topic size and the fuzziness value impacts the error rates. For all measures, the error rate decreases as more topics are used. They also show that when the fuzziness value is increased, the error rate decreases but the discriminative power is reduced since more measures are considered equal and therefore fewer conclusions can be drawn.

5.1.2. *Achieved significance level*

Sakai (2006) uses a bootstrapping approach for calculating the sensitivity of evaluation measures. It is based on the same underlying idea of quantifying the differences of performance distributions and counting the number of times an evaluation measure produced significantly different performance distributions based on varying query sets. The main difference to the error rate approach is that the bootstrap query sets are created by sampling with replacement from the set of topics.

Sakai (2006) defines an *achieved significance level* (ASL) using a paired bootstrap hypothesis test. Let *Q* be the set of topics, where $|Q| = n$. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be the vectors containing the per-topic performance values of system X and Y as computed by some evaluation measure *M*. Since **x** and **y** is paired data, let $\mathbf{z} = (x_1 - y_1, \ldots, x_n - y_n)$ since we are interested in the difference in population means for *X* and *Y* based on the population *P* of topics. Therefore, let $\mu = \mu_X - \mu_Y$ with the following hypotheses for a two-tailed test:

$$H_0 : \mu = 0 \quad vs. \quad H_1 : \mu \neq 0. \tag{14}$$

Thus it is assumed that **z** is an independent and identically distributed sample drawn from an unknown distribution. Let *t* be the test statistic with the following null hypothesis distribution:

$$t(\mathbf{z}) = \frac{\bar{z}}{\bar{\sigma}/\sqrt{n}} \tag{15}$$

where $\bar{\sigma}$ is the standard deviation of **z**.

Moreover, let $\mathbf{w} = (w_1, \ldots, w_n)$ where $w_i = z_i - \bar{z}$ is a bootstrap sample of the per-topic differences that obeys H_0. Consider an example similar to Table 2 but with a topic size $n = 5$ where $Q = (T_1, T_2, T_3, T_4, T_5)$ and $\mathbf{w} = (0.2, 0.1, 0.8, 0.3, 0.1)$. For a bootstrap sample $b = (T_2, T_1, T_1, T_5, T_3)$, the distribution of differences is $\mathbf{w}_b = (0.1, 0.2, 0.2, 0.1, 0.8)$. The ASL is computed by sampling *B* bootstrap samples with replacement and computing the rate of the number of times that $|(\mathbf{w}_b)| \geq |(\mathbf{z})|$. The rate indicates how rare the observed difference would be under H_0. For a significance level of α, H_0 can be rejected if ASL < α. In other words, μ_X and μ_Y are different with a certain probability.

In summary, for each system pair *X*, *Y* and an evaluation measure *M*, the rate of the number of times that ASL < α is computed based on *B* bootstrap samples. The more frequently an evaluation measures' ASL falls below α, the more sensitive it is to detect differences between lists of rankings. For example, given an ASL < 0.05 and a rate of 80% for measure *M*, we can be 95% confident that *M* identifies differences in rankings 80% of the time.

[2] http://trec.nist.gov/data/intro_eng.html.

Table 4
A depiction of a test collection of relevant author entities. Author entities are split into 25 groups such that one relevant entity $\text{Author}_{i,t}$ from group t belongs to exactly one query Q_i. In this experimental design no distinction between query types exists.

Query Type	Group 1	Group 2	…	Group 25	Query
1	$\text{Author}_{1,1}$	$\text{Author}_{1,2}$	…	$\text{Author}_{1,25}$	$\mathcal{Q}_1$
	$\text{Author}_{2,1}$	$\text{Author}_{2,2}$	…	$\text{Author}_{2,25}$	$\mathcal{Q}_2$
	⋮	⋮	⋱	⋮	⋮
	$\text{Author}_{10,1}$	$\text{Author}_{10,2}$	…	$\text{Author}_{10,25}$	$\mathcal{Q}_{10}$

The same approach is used by Shi, Tan, Zhu, and Wu (2013) where they use the two-sample t-test for equal sample sizes and equal variances. Again, the result lists of two systems over a query set is compared using the t-test at significance level 0.05. The sensitivity rate for an evaluation measure is the percentage of pairs for which a significant difference in the result distributions is observed. They find that nDCG performs the best, followed by AP and P@10 with sensitivity rates ranging from 35% to 85%.

5.1.3. Swap rate

Voorhees and Buckley (2002) initially proposed the swap method for estimating the difference required by a measure when varying the significance levels and topic sizes. In this paper, we focus on the bootstrapped approach used by Sakai (2006) where we sample with replacement. The idea behind the swap method is to estimate the "swap rate", which gives the probability that two experiments are contradictory given an overall performance difference. For each evaluation measure M, a list of 21 "performance difference bins" are created which is used to record the performance differences between two systems (Voorhees & Buckley, 2002). Let D be the performance difference between two system X and Y as measured by M. The first bin represents performance differences such that $0 \leq D < 0.01$, the second bin represents those such that $0.01 \leq D < 0.02$, until the last bin which represents those such that $0.20 \leq D$. Two bootstrap samples (Q_{b1}, Q_{b2}) are selected from the set of queries and two performance differences calculated between two systems (X and Y). For example, $D_{b1} = M(X, Q_{b1}) - M(Y, Q_{b1})$ and $D_{b2} = M(X, Q_{b2}) - M(Y, Q_{b2})$. If the order in terms of performance of the two systems ($D_{b1} \cdot D_{b2} \leq 0$) changes, the swap counter in the associated bin (D_{b_1}) is increased. A second counter is associated with each bin which simply keeps track of the number of times an associated performance difference occurred. The swap rate of a performance difference interval is simply the swap counter in the associated bin divided by the total number of occurred performance differences of that bin.

By iterating through the performance difference bins and finding the bin where the swap rate drops below a certain significance level (say 0.05) one can estimate how much difference is required in order to conclude that system X is better than Y with 95% "confidence". It should be noted that this method is not directly related to statistical significance testing.

5.2. Experimental framework for rankings of academic entities

In the case where evaluation measures are evaluated on ranks of academic entities the above described experimental methods have to be adapted for various reasons. In this context, the document set is the complete set of entities (e.g., all authors or all papers) in a database and are not categorised into topics.

The notion of different user information needs (i.e., topics) can be mimicked by simply constructing different sets of relevant entities. This is possible since no distinction is made between relevant entities and they can therefore be randomly placed into different groups. It should be noted that we have no way to construct different queries to try and retrieve the same group of relevant entities. Therefore, we use different groups of relevant entities as query sets to observe evaluation behaviour changes and to calculate associated error, tie, swap, and ASL rates.

Assume we divide a set of r relevant entities into 25 groups. We obtain 25 queries each consisting of $r/25$ relevant entities. The 25 queries are created by sampling without replacement from the r relevant entities. Therefore each relevant entity is used exactly once. For the ASL and swap rates, the bootstrap samples are constructed by sampling with replacement from the set of 25 queries and therefore relevant entities can be used more than once (Table 4).

Table 5 shows the decision table of two evaluation measures (average and median rank) using three author ranking systems (citation counts, h-index, and author-level Eigenfactor West, Jensen, Dandrea, Gordon, & Bergstrom, 2013) with 25 queries each consisting of 37 relevant entities (ACM fellows).

In this example the error and tie rates of the average rank measure are 2.667% and 5.333%, while the corresponding rates of the median rank measure are 40.000% and 10.667%. The ASL rate ($\alpha < 0.05$) for the average rank measure is 100%. This indicates that the average rank measure can distinguish between rankings of the three systems every time and requires a difference of 290.253 in scores. The average rank of the systems are 3832.08 (Eigenfactor), 7008.33 (Citations), and 7582.07 (h-index). In this case we can conclude with 95% "confidence" that the Eigenfactor measure is better than using citations, which in turn is better than the h-index for identifying ACM fellows.

Table 5
Illustrative example of counts of the number of times the system of the row was better than, equal to, or worse than the system of the column. A table is shown for average rank and median rank measure. The error rate and tie rate are given based on the associated decision tables. The ASL rate ($\alpha < 0.05$) and estimated difference (Est. Diff.) are computed with 1000 bootstrap samples.

Average	*h*-Index			Eigenfactor		
Citations	23	1	1	0	3	22
h-Index				1	0	24
Error rate:	2.667%					
Tie rate:	5.333%					
ASL rate:	100%					
Est. Diff.	290.253					

Median	h-index			Eigenfactor		
Citations	10	3	12	12	4	9
h-Index				13	1	11
Error rate:	40.000%					
Tie rate:	10.667%					
ASL rate:	0%					
Est. Diff.	74.289					

As mentioned before, rankings of academic entities do not depend on the number of topics or relevant entities compared to. However, the number of entities per query should be kept the same since some evaluation methods depend on the number of relevant entities.

We use the error and tie rates to judge the performance of the various evaluation measures. We use the bootstrap approach proposed by Sakai (2006) to compute a sensitivity rate for measures, and adapt the swap method using relative differences to estimate performance differences required to judge results of a system X significantly different to system Y.

Since not all evaluation measures discussed in this paper produce scores that range between 0.0 and 1.0, instead of static performance difference bins, we use relative performance difference bins. Let min_M and max_M be respectively the minimum and maximum difference obtained using evaluation measure M. The bins are defined as follows

$$\text{Bins}_M = \{\text{Bin}_{M_1}, \text{Bin}_{M_2}, \ldots, \text{Bin}_{M_{21}}\} \tag{16}$$

where

$$\text{Bin}_{M_i} = \left[\min_M + \frac{i-1}{5} \cdot \frac{\max_M - \min_M}{21}, \min_M + \frac{i}{5} \cdot \frac{\max_M - \min_M}{21}\right) \tag{17}$$

The same number of bins (21) are used as in previous work (Sakai, 2006; Voorhees & Buckley, 2002) as well as the last bin having its boundary at 20% of the maximum score difference.

6. Results

We use the framework described in the previous section to analyse the performance of evaluation measures for evaluating sparse rankings of academic entities. We use author (ACM fellows and LCA authors) and paper entities (BPA papers and HI papers) as separate sets of relevant entities on both the ACM and MAG databases. These sets of relevant entities have different citation properties. For example, the BPA papers have fewer citations than HI papers. On the ACM database the average number of citations for BPA and HI papers are 43.05 and 229.49. Furthermore, the ACM and MAG database sizes differ substantially. For example, the relevant author rankings are evaluated on 32 million MAG author entities, compared to 625 566 author entities using the ACM database. The goal of using different test data sets and real-world databases with varying properties is to gain insight into the stability of the evaluation measures and their general appropriateness.

We use the fractional ranking approach described in Section 2 to convert scores into ranks. Precision values greater than 1 are adjusted downwards to 1 as indicated in Section 4.2. For permille ranks we use the approach by Waltman and Schreiber (2013) to compute fair ranks across permille boundaries and the methodology described in Section 4.1 to compute the evaluation measure's scores.

Table 6 shows the evaluation results based on eight author ranking algorithms computed on the ACM database. The test data comprises the 930 ACM fellows which are split into 15 groups with 62 relevant entities per group. The ranking algorithms used here and in the following sections (unless stated otherwise) are citation counts, the author-level Eigenfactor metric, *h*-index, *g*-index, standard PageRank, publication counts, *PR*-index and co-author counts. For each algorithm all self-citations are included. We selected these algorithms since they have relatively different ranking properties with different rank distributions. The overall average Spearman rank correlation coefficient between each pair of ranking algorithms is 0.720 ($\sigma = 0.181$).

All evaluation measures listed in the table are described in Section 4. We use different cut-offs for measures that define cut-off thresholds but only show the results of the best performing ones according to the ASL rate. The considered cut-off thresholds are 10, 100, 2000, the number of relevant entities in a query, the average rank of the relevant entities in a query

Table 6
The results of evaluating the evaluation measures using eight ranking algorithms. The ACM fellows are used as relevant entities and split into 15 queries against the ACM database. All results are averaged over 50 iterations. A significance level of 0.05 is used for the ASL rate with 1000 bootstrap samples. The error and tie rates respectively indicate the measures' lack of stability and lack of discriminative power. The column "Rel" shows the average percentage of relevant entities used by the measures. The column "Est. Diff" shows the minimum difference in scores required to consider the performance of two ranking algorithms significantly different. Based on the estimated required difference, column "Tot. Diff." shows the percentage of ranking algorithm pairs for which a measure detected significant differences in rankings.

Measure	ASL (%)	Error (%)	Ties (%)	Rel (%)	Est. Diff.	Tot. Diff. (%)
P@Avg	93.500	3.657	4.843	79.998	0.001	100.000
ROC	91.286	0.124	28.919	84.234	0.016	89.286
nDCG@Avg	90.429	3.990	5.610	78.387	0.001	100.000
Average	90.000	3.914	6.481	100.000	985.069	89.286
Average@Avg	80.000	7.429	12.567	78.387	253.837	60.714
Recall@2000	64.571	9.200	19.919	55.954	0.017	75.000
ROC@2000	64.286	9.414	15.867	84.234	0.015	89.286
PR	59.786	13.743	10.790	100.000	0.003	96.429
AP@0.5recall	53.857	15.881	9.186	50.844	0.003	96.429
Median@0.5recall	51.929	15.910	11.676	50.844	118.835	46.429
Max	49.643	17.833	6.071	1.613	35612.024	78.571
Median	49.071	14.448	15.810	100.000	318.926	46.429
nDCG	44.786	5.481	47.500	100.000	0.011	60.714
AP	43.500	16.281	11.138	100.000	0.004	96.429
Min	7.286	35.395	6.467	1.613	10.534	0.000

(i.e., P@Avg), and the rank at which 0.5 recall is achieved (i.e., AP@0.5recall). The threshold intervals used by ROC are [100, 200, . . ., 10 000].

All results are average rates based on 50 iterations where the ACM fellows are randomly split into different groups for each iteration. The evaluation measures in the table are sorted in descending order of the ASL rate with $\alpha = 0.05$. The fuzziness value for computing the error and tie rates is 5%. The column "Rel (%)" lists the average percentage of relevant entities used by the evaluation measures for computing the scores. For example, the Max measure only uses a single relevant entity for each query which is the rank of the entity with the largest rank value. Since the results are based on 15 queries, 1.613% of the 930 ACM fellows are used for computing Max scores for each iteration.

The column "Est. Diff." shows the difference in performance scores required by a measure to conclude that two systems rank the ACM fellows significantly different (guaranteeing a swap rate of 5% or less). For example, when using the average rank measure, a difference of 985.1 in the average rank would be required to make any judgements with 95% confidence. Note that these values can not be used to compare evaluation measures to each other because of the measures' differing score ranges.

The column "Tot. Diff. (%)" shows the overall differentiation rate of the evaluation measures when the estimated required difference is used as a threshold to differentiate between the ranking algorithms using the complete set of relevant entities in a single query. In other words, it shows the percentage of ranking algorithms pairs for which the evaluation measure differentiated the rankings with 95% confidence.

A high error rate indicates a lack of stability. The Min and Max measures have the highest error rates with 35.395 and 17.833, respectively. This is not surprising since they only use the ranking of one relevant entity per query to compute performance scores.

The AP and nDCG measures should be highlighted since they both sum performance values over each rank of the relevant entities but with different score distributions. AP gives more weight to higher ranked entities and the utility of lower ranked entities tends towards zero. The score distribution of nDCG is long-tailed where the difference in utility of higher ranked entities is smaller compared to AP. Furthermore, the relevant entities at lower ranks have more weight in the overall score. Both measures have low sensitivity with ASL values of 43.500 and 44.786. The low sensitivity of AP is due to relatively high instability (error rate of 16.281) and average discriminative power (tie rate of 11.138). For nDCG this is switched around: it is relatively stable with an error rate of 5.481, but has very low discriminative power with a tie rate of 47.500.

It should be noted that for all measures that use at least half of the relevant entities, the tie rate is relatively low except for nDCG (47.500) and ROC (28.919). We assume that the relevant entities that are ranked very low, and have very similar scores independent of their actual rank, dominate the overall result of the nDCG measure. This is supported by the low tie rate (5.610) of nDCG@Avg which uses a cut-off threshold at the average rank of the relevant entities. In other words, the discriminative power of nDCG is improved substantially by removing the relevant entities with lower than average rank.

For the ROC measure a similar argument applies. We assume that the high false positive rate at large cut-off threshold values dominates the overall value of the area under the curve. An indication of this is that when ROC is calculated at ranks defined by the relevant entities and with a cut-off at 2 000 (ROC@2000), the tie rate drops significantly to about 15.867.

The top three measures according to the ASL rates are P@Avg (93.500), ROC (91.286), and nDCG@Avg (90.429). However, ROC has relatively low discriminative power with a tie rate of 28.919. When considering the error and tie rates together, the only measures that have both rates below 10% are P@Avg, nDCG@Avg, and the Average measure.

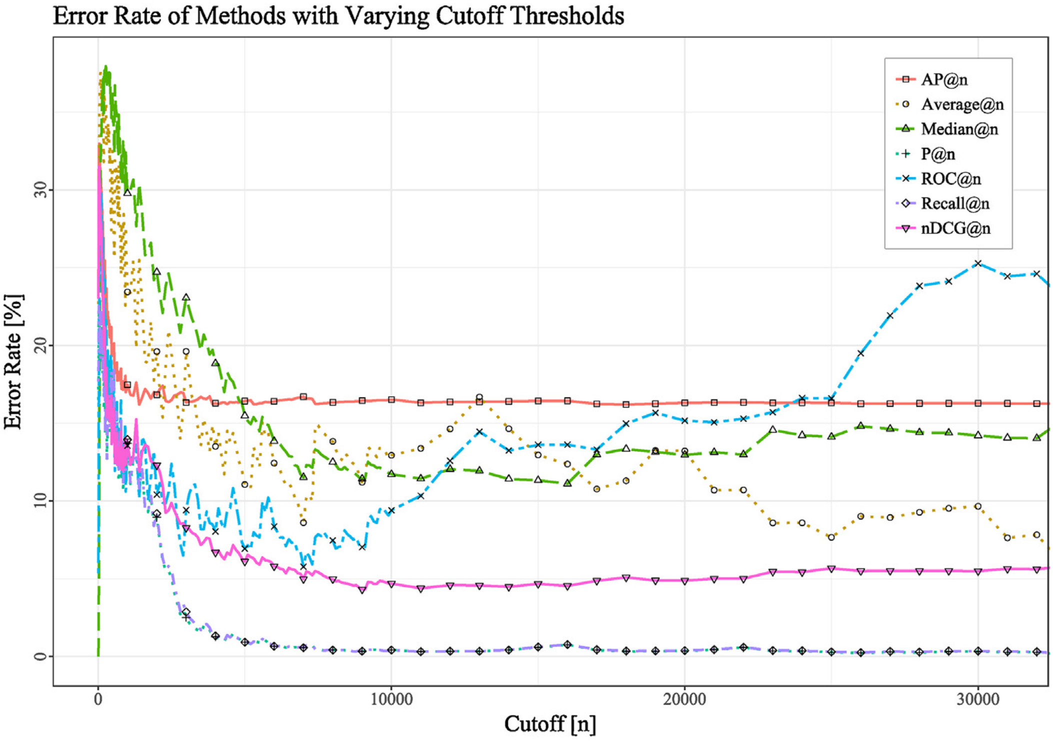


Fig. 7. The average error rates of evaluation methods that define cut-off thresholds with varying cut-off values. For all measures the query size is fixed at 15 with the fuzziness value set to 5%.

As we assumed before, the error and tie rates of most evaluation measures are very high when compared to the results of query-based information retrieval systems highlighted in Section 5.1. Furthermore, we mentioned that the median rank has less discriminative power than the average rank measure which, in this case, is also true where the respective ASL rates are 49.071 and 90.000. Consequently, the overall differentiation rate of the average rank measure (89.286) over all ACM fellows is also higher compared to the median rank measure (46.429).

The results in Table 6 only give an indication of the performance of the evaluation measures. The values are expected to change when the cut-off thresholds of the measures change, the parameters of the framework are varied, and different data sets are used. We show this in the following subsections.

6.1. Varying the cut-off thresholds

To better understand the stability of the evaluation measures that define cut-off thresholds, we compute the error and tie rates while varying the cut-off threshold. The ACM fellows on the ACM database are used with a fuzziness value of 5%, 15 queries, and 50 iterations over which the results are averaged. Fig. 7 shows the average error rates of the evaluation measures when the cut-off threshold is varied. The error rate has to be considered together with the corresponding tie rates shown in Fig. 8.

Since an average error rate above 25% is considered to be dominated by noise (Buckley & Voorhees, 2000), we argue that if the error rate of a metric does not drop significantly below 25% it has no utility for evaluating the ranks under consideration. Similarly for the tie rate; if a metric's tie rate is above 50%, we argue that its discriminative power has no utility since it cannot discern the better of two rankings more than half the time.

Fig. 7 shows that the error rates of all measures are very high for small cut-off values. All error rates decrease when the cut-off threshold is increased and level out after certain cut-off values. The exception is ROC@*n* which increases substantially for cut-off values above 6000. The error rates of Average@*n* and Median@*n* only fall below 15% at cut-off thresholds of 3400 and about 5700, respectively. The error rate of AP does not fall below 15% but has a tie rate relatively stable at 11%. The measures with the lowest error rates on average are nDCG@*n*, Recall@*n* and P@*n* for all cut-off values.

Even though these measures have the lowest error rates, their discriminative power is the least, with tie rates above 25% after a cut-off threshold of 5000. Also, as Fig. 8 shows, their tie rates increase steadily. This is expected since the value of Recall@*n* only increases with the cut-off threshold until it reaches 1 when all relevant entities are retrieved, at which point the tie rate is 100%. The score of P@*n* decreases and tends towards 0 with larger cut-off thresholds since the number of non-relevant entities increases. Therefore, P@*n* will inevitably reach tie rates of 100% with large enough cut-off values. For nDCG@*n* the steady increase in the tie rate can be explained by the aforementioned long-tailed distribution. Except for these three measures, all other tie rates level off and stay stable for cut-off values above certain values. In general, the lowest tie rates for all measures are at cut-off thresholds between 200 and 1000.

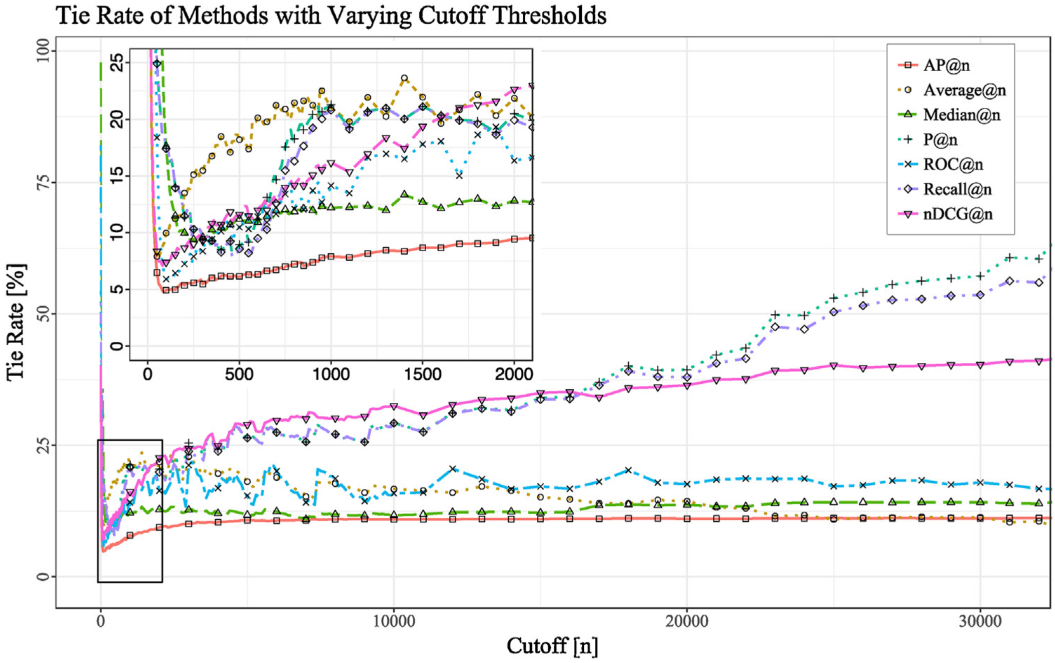


Fig. 8. The average tie rates for the evaluation methods across varying cut-off values corresponding to the error rates shown in Fig. 7. Only evaluation measures that define a cut-off threshold are shown.

Surprisingly, Median@*n* (tie rate ≈ 12%) has a higher discriminative power than Average@*n* (tie rate ≈ 16%). However, from Table 6 we know that Average (error rate = 3.914, tie rate = 6.481) outperforms Median (error rate = 14.448, tie rate = 15.810) when no cut-off value is used and therefore the two tie rate curves have to switch order. In this case the switch occurs at *n*= 20 000.

In conclusion, AP@*n* seems to be the most stable under varying cut-off thresholds. nDCG@*n*, Recall@*n* and P@*n* have the lowest error rates but their discriminative power decreases dramatically when the cut-off threshold value is too large and therefore appropriate cut-off thresholds should be used. Similarly, ROC@*n* also requires an appropriate cut-off value since its error rate increases for larger cut-off values.

6.2. Varying the significance levels

The experimental design uses a notion of relative difference to consider the distributions of two rankings equal. The method proposed by Buckley and Voorhees (2000) defines a fuzziness value which uses the variance (absolute difference) in the mean of evaluation scores. The higher this significance level the more likely two distributions of scores are considered equal. By relaxing the significance level (increasing the fuzziness value) the error rates should naturally decrease while the tie rates should increase. Analysing the stability of the evaluation measures over a range of significance levels is crucial since a fixed fuzziness value could imply different trade-offs for different measures.

Fig. 9 shows the error rates plotted against the tie rates of the measures when the significance value is varied between 0.005 and 0.1 with intervals of 0.005. For evaluation measures that define cut-off thresholds we only show the results for which the best performance is obtained. The measures Average@*n*, P@*n* and nDCG@*n* perform the best when the cut-off threshold is set to the average rank of the relevant entities. For the AP@*n* and Median@*n* the best results are obtained when the cut-off is set to the rank at which 50% recall is achieved. Lastly, Recall@*n* performs the best with a cut-off value of 2000.

As expected, the error rates decrease with higher significance values since the threshold of differences between distributions is relaxed. Similarly, the tie rates increase when the significance value is increased. The measures P@Avg, nDCG@Avg, Average and Min vary the least. However, Min has a very high error rate. It should be noted that varying the significance levels in the framework does not advantage or disadvantage any measure disproportionately. This is important since it means that the experimenter can unreservedly vary the significance level. For example, the experimenter can relax the significance level by choosing a larger significance value should it be required to distinguish between more ranking algorithms. The trade-off is a decrease in the confidence level about the conclusions drawn from those results.

6.3. Varying the query set size

Fig. 10 shows the average error and tie rates of the evaluation measures and how they change when the query sizes are varied. For a query size of 3, the entire test set (ACM fellows) is split into three queries (against the ACM database). The larger

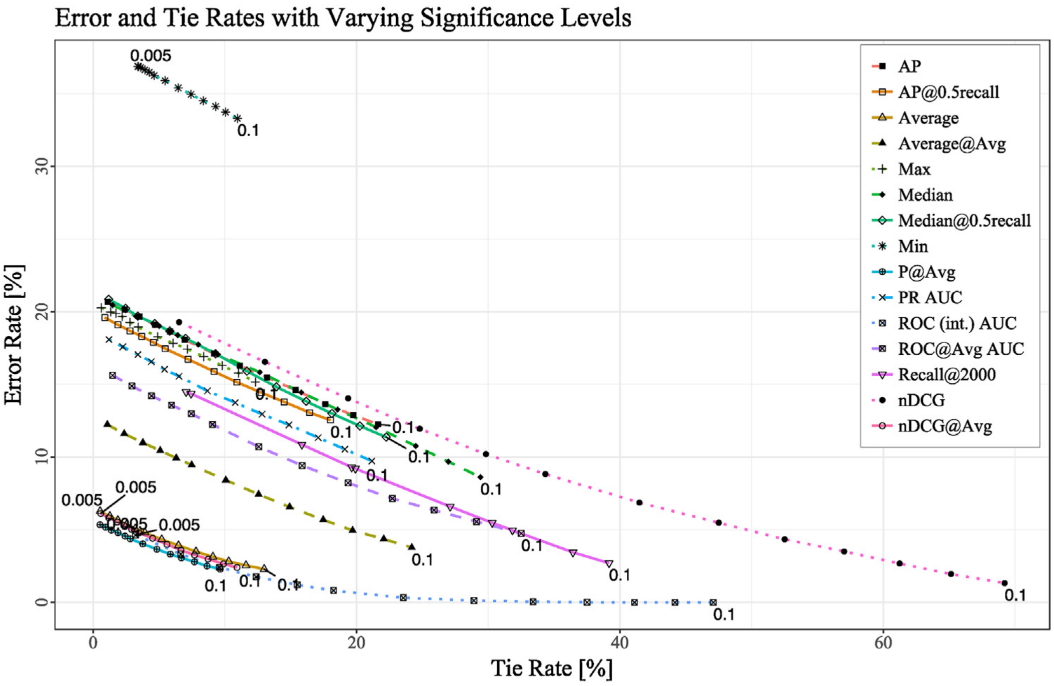


Fig. 9. The error and tie rates of the evaluation measures with varying significance levels. The significance value is varied between 0.005 and 0.1 with 0.005 intervals. For evaluation measures that define cut-off thresholds only the best performing variation is shown.

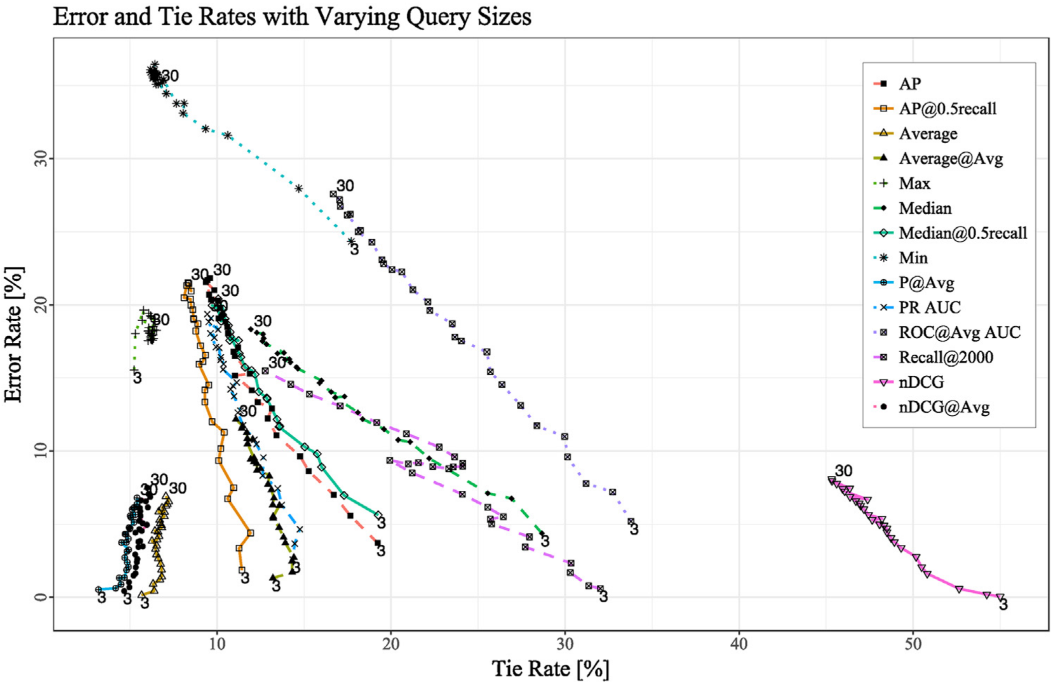


Fig. 10. The average error and tie rates for the evaluation measures over varying query sizes of ACM fellows computed on the ACM database.

the query size, the smaller the number or relevant entities per query but the more decisions are made when computing the error and tie rates. The significance level is set to 0.05 with 50 iterations over which the results are averaged.

From Fig. 10 one can see that the error rates increase with the query size. This is expected since the evaluation measures have less information to work with. For all methods except P@Avg, nDCG@Avg, Average and Max, the tie rates decrease steadily when the query sizes are increased. Here, it is important to look at the variance in the error and tie rates since it indicates how stable the evaluation measures are when less information is available.

The worst performing measures are Min, ROC@Avg, and nDCG since they have either high error or tie rates. Furthermore, Median and Recall@2000 have relatively low discriminative power for small query sizes. It is unclear why the tie rate of Recall@2000 changes directions twice at a query size of 15 and again at 21. The most stable measures are P@Avg, nDCG@Avg, Average and Max, but Max has a high error rate.

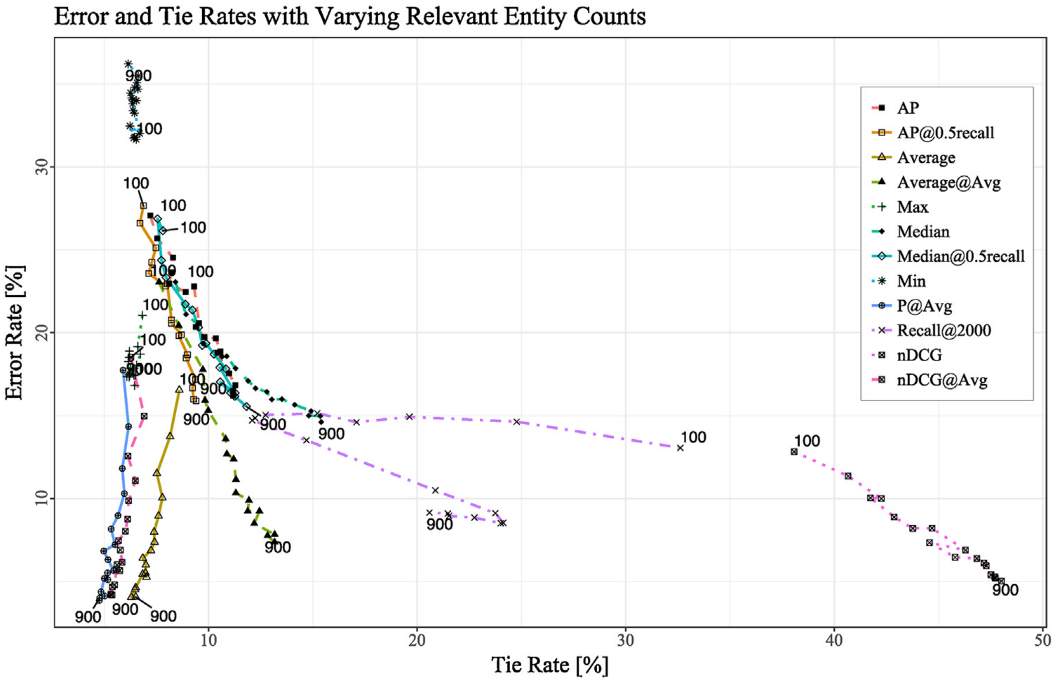


Fig. 11. The average error and tie rates of evaluation measures with varying relevant entity counts.

6.4. Varying the relevant entity counts

Fig. 11 shows the average error and tie rates of the evaluation measures when different number of relevant entities (ACM Fellows) are used. The entity count is varied between 100 and 900 ACM fellows (intervals of 50) which are always split into 15 groups. Therefore, each query size varies between 6 and 60 relevant entities. The other parameters are kept constant with a significance value of 0.05.

In general, most measures' error rates decrease when more relevant entities are used. The only exception is the Min measure, whose error rate is the highest with the most relevant entities. The tie rates of the measures vary less compared to the error rates, except for the nDCG and Recall@2000 measures. Again, it is unclear why the Recall@2000 measure is unstable when the underlying data is varied.

The tie rate increases with more relevant entities except for P@Avg, nDCG@Avg, Average and Max. Again, the most stable measures with low error and tie rates are P@Avg, nDCG@Avg and Average.

6.5. Varying the databases

All results discussed up to this point are based only on the ACM database and a single set of relevant entities. Table 7 shows the top 3 evaluation measures, based on their ASL rates, for each test data set on both the ACM and MAG databases. In addition, we show the results of the Average and Median measures since we are interested in their performance differences and whether the Average measure is appropriate in general. Furthermore, we show the results of nDCG@Avg since it showed promising results in the previous sections and contrast it to the results of nDCG without a cut-off.

The test data sets are split into 15 groups and the significance level is set to 0.05. Since it is expected that the results vary between databases when different cut-off thresholds are used, we use values that are not static but based on the data. As cut-off thresholds we therefore use the average rank of the relevant entities and the rank at which 50% recall is achieved.

For each measure, Table 7 shows the ASL, error and tie rates, as well as the estimated difference required in order to judge scores between two ranking algorithms different ($\alpha = 0.05$). Furthermore, the column 'Tot. Diff.' shows a measure's actual rate of differentiation between ranking algorithms. In other words, when comparing all ranking algorithm pairs on the complete test data set, 'Tot. Diff.' shows the percentage of comparisons for which a measure could identify a significant difference in the rankings based on the estimated required difference and the chosen significance level.

One can clearly observe from the results that there does not exist any obvious best measure. The nDCG measures without a cut-off is always in the top three when using the MAG database. However, it has low discriminative power (high tie rates) as observed in the previous sections. The discriminative power is improved when nDCG uses a cut-off at the average rank of the relevant entities. When using paper test data (BPA and HI) on the MAG database, the average rank measure performs the best. For the ACM database, the nDCG@Avg and P@Avg appear in the top three for three of the four test data sets. The exception is for BPA where nDCG has the highest sensitivity according to the ASL rate. However, it has an extremely high tie rate of 75.014%.

Table 7
The top tree evaluation measures (based on the ASL rate with $\alpha = 0.05$) for each test data set and publication database. In addition, the results of the Average, Median, nDCG and nDCG@Avg measures are shown for comparison. Column 'Est. Diff.' shows the estimated difference required when comparing two ranking algorithms. The actual percentage of differentiation achieved by a measure between all pairs of ranking algorithms based on the estimated required differences is given in column "Tot. Diff.".

Database	Test Data	Measure	ASL $< \alpha$	Error (%)	Tie (%)	Est. Diff.	Tot. Diff.
ACM	ACM	P@Avg	93.500	3.657	4.843	0.00089	100.00
		ROC	91.286	0.124	28.919	0.01550	89.286
		nDCG@Avg	90.429	3.990	5.610	0.00107	100.00
		Average	90.000	3.914	6.481	985.06857	89.286
		Median	49.071	14.448	15.810	318.92583	46.429
		nDCG	44.786	5.481	47.500	0.01064	60.714
	LCA	ROC	93.643	0.738	20.386	0.02195	89.286
		P@Avg	85.786	7.319	5.671	0.00046	96.429
		nDCG@Avg	85.286	7.133	5.333	0.00055	96.429
		Average	70.929	8.662	7.533	2415.11273	78.571
		Median	63.214	10.605	10.419	524.24046	46.429
		nDCG	49.500	8.852	33.252	0.01405	64.286
	BPA	nDCG	91.857	0.000	75.014	0.00136	96.429
		Average@0.5.recall	86.786	5.338	6.448	2935.20399	89.286
		PR	83.214	5.510	6.843	0.00003	100.000
		nDCG@Avg	82.643	7.438	8.548	0.00001	100.000
		Median	81.143	8.476	6.529	8847.73274	85.714
		Average	78.929	8.690	9.562	9059.78994	89.286
	HI	Average	74.143	12.810	8.762	5116.65704	67.857
		nDCG@Avg	71.857	11.367	6.843	0.00015	100.000
		P@Avg	69.214	12.510	6.776	0.00013	100.000
		nDCG	60.143	2.167	52.295	0.00700	71.429
		Median	53.714	15.919	12.657	2046.40451	28.571
MAG	ACM	nDCG	98.000	0.000	16.690	0.00383	100.000
		ROC	97.500	1.767	2.324	0.01061	100.000
		P@Avg	97.143	1.495	3.581	0.00003	100.000
		nDCG@Avg	96.714	1.281	3.076	0.00003	100.000
		Median	95.857	0.429	8.624	9191.87913	85.714
		Average	93.071	2.048	4.448	49058.26690	85.714
	LCA	Average@0.5.recall	97.857	0.048	13.952	6303.55246	78.571
		ROC	96.571	1.800	2.929	0.01424	100.000
		nDCG	96.500	0.062	13.086	0.00428	100.000
		Median	92.500	0.595	10.967	13631.44058	82.143
		Average	70.143	11.667	8.305	186875.15184	64.286
		nDCG@Avg	62.429	6.462	5.243	0.00001	96.429
	BPA	Average	99.857	0.667	7.043	150938.26963	100.000
		P@Avg	98.143	1.310	12.581	0.00000	100.000
		nDCG	97.071	0.000	60.014	0.00076	96.429
		nDCG@Avg	96.571	1.352	11.524	0.00000	100.000
		Median	90.357	5.219	5.662	317670.01455	92.857
	HI	Average	90.857	5.995	8.138	97070.09494	89.286
		nDCG	87.143	0.048	40.467	0.00175	89.286
		Average@0.5.recall	80.714	6.095	9.771	8129.54409	75.000
		nDCG@Avg	74.429	4.743	9.290	0.00000	96.429
		Median	66.286	10.762	8.010	33108.36986	67.857

The Average rank measure has higher discriminative power (smaller tie rate) than Median in almost all cases. The exception is when the BPA test data set is used. On the MAG database Median also has higher sensitivity (ASL rate) for the ACM fellows and LCA authors. Only for the LCA authors on the MAG database does Median perform significantly better than Average.

Lastly, the differentiation rates (Tot. Diff.) are relatively high (ranging between 67% and 100%) which indicates that these measures can successfully identify significant differences in the rankings most of the time independent of the database or relevant entities used. The lowest ASL value for the average rank measure is 70.143 when the LCA authors are used on the MAG database. Nonetheless, the error and tie rates of Average are consistently low. This shows that Average is a

Table 8
The top tree evaluation measures (based on the ASL rate with $\alpha = 0.05$) for each test data set and publication database computed on permille rankings. For comparison the results of Average, Median, nDCG and nDCG@Avg are also given. Column 'Est. Diff.' shows the estimated difference required when comparing two ranking algorithms. The actual percentage of differentiation achieved by a measure between all pairs of ranking algorithms based on the estimated required differences is given in column "Tot. Diff.".

Database	Test Data	Measure	ASL $< \alpha$	Error (%)	Tie (%)	Est. Diff.	Tot. Diff.
ACM	ACM	P@Avg	93.500	3.776	4.914	0.00094	100.000
		AP@Avg	91.929	3.295	6.543	0.00140	96.429
		Average	89.929	3.819	6.810	1.61272	85.714
		nDCG	61.214	2.729	39.838	0.02003	53.571
		Median	48.500	9.867	26.076	0.46054	57.143
		nDCG@Avg	48.286	7.167	35.243	0.02142	46.429
	LCA	AP	98.214	0.014	33.624	0.00000	100.000
		nDCG	91.643	2.524	21.638	0.02284	89.286
		AP@Avg	87.214	4.338	6.019	0.00070	100.000
		nDCG@Avg	79.786	5.805	20.129	0.02655	85.714
		Average	70.286	8.576	7.805	3.80409	78.571
		Median	63.786	8.043	17.833	1.04334	46.429
	BPA	AP	95.786	0.838	21.043	0.00000	100.000
		AP@Avg	95.714	2.790	6.571	0.00002	100.000
		AP@0.5recall	92.571	3.233	6.281	0.00002	100.000
		nDCG	91.000	1.471	23.129	0.00371	89.286
		nDCG@Avg	88.929	3.019	15.214	0.00566	92.857
		Median	84.214	7.267	7.048	8.25132	89.286
		Average	83.643	7.986	14.790	6.49719	89.286
	HI	AP@Avg	85.500	5.648	7.886	0.00020	100.000
		Average	85.214	8.443	12.933	3.63897	78.571
		AP	85.000	0.110	43.062	0.00000	100.000
		nDCG	84.429	1.752	25.414	0.01788	78.571
		Median	84.214	7.267	7.048	8.25132	89.286
		nDCG@Avg	76.143	3.867	22.938	0.01919	75.000
MAG	ACM	AP	100.000	0.000	18.181	0.00000	100.000
		nDCG	98.143	0.048	6.462	0.01777	96.429
		nDCG@Avg	97.000	0.557	7.171	0.02044	96.429
		Average	93.357	2.024	4.848	1.55892	85.714
		Median	90.929	0.690	15.410	0.97691	75.000
	LCA	AP	100.000	0.000	16.062	0.00000	100.000
		nDCG	100.000	0.067	5.724	0.01398	100.000
		nDCG@Avg	99.786	0.300	5.500	0.01540	100.000
		Median	86.357	0.776	15.695	1.16592	75.000
		Average	69.000	11.710	8.600	5.70714	64.286
	BPA	Average	99.857	0.671	7.062	7.90347	100.000
		ROC	99.857	0.014	29.976	0.00829	100.000
		nDCG	98.857	0.743	18.686	0.00348	96.429
		nDCG@Avg	92.357	2.181	12.757	0.00647	96.429
		Median	86.357	0.776	15.695	1.16592	75.000
	HI	Average	90.429	5.967	8.219	5.00508	89.286
		nDCG@Avg	85.143	3.448	17.371	0.01570	82.143
		nDCG	83.000	3.138	19.676	0.01574	82.143
		Median	65.286	10.524	9.581	1.05064	67.857

stable measure with high sensitivity and discriminative power. It should also be noted that the performance of Average can sometimes be improved by using a cut-off threshold at 50% recall.

Table 8 is similar to Table 7 except that the permille based rankings are used. Again, the results are very noisy. However, a few observations are possible. The nDCG measure is again in the top three for each test data set using the MAG database. It should be noted that using permille rankings, the discriminative power of nDCG improves significantly. This is expected since the maximum rank value of relevant entities is 1000, limiting the impact of nDCG's long-tailed rank utility distribution. Another indication of this is that the discriminative power of nDCG@Avg only slightly improves over nDCG without a cut-off.

The AP measure both with and without cut-off thresholds is frequently in the top three which is not the case when normal rankings are used. However, the discriminative power of AP is very low but is increased when a cut-off value is used.

Similar to the normal rankings, the Average rank measure has higher discriminative power than Median in all except one case (BPA on the ACM database). In addition, the Median measures has higher sensitivity for the LCA test data set on the MAG database. Otherwise, Average performs better than Median. Since the Average and Median measures are computed the same way using permille rankings as for normal rankings, it is not surprising that they produce very similar results. For both measures the ASL rates are very similar compared to when normal rankings are used. Furthermore, the error and tie rates of the Average measure only deviate by 0.400% except for the BPA and HI data sets on the ACM database. However, for the Median measure the discriminative power is reduced when permille rankings are used.

6.6. *Varying ranking algorithms*

Up to this point, we used 8 different ranking algorithms for ranking both authors and papers. In this section we show results of when the number of ranking algorithms is varied. The rankings are chosen randomly from a set of 64 author and 38 papers ranking algorithms, all of which have been described in previously published work (Dunaiski et al., 2016, 2018; Nykl et al., 2015). Due to space reasons we do not list the algorithms. Some rankings are very similar where only a parameter of an algorithm is changed. For example, excluding or including self-citations for papers and authors or changing the damping factor of the PageRank algorithm. Other rankings differ more such as ranking authors according to their co-author counts versus the rankings produced by the author-level Eigenfactor metric. The minimum and maximum average Spearman rank correlation coefficient between the sets of rankings used in this analysis is 0.638 and 0.953.

Fig. 12 shows the error and tie rates of the evaluation measures when varying the number of rankings used. For each iteration the rankings are randomly drawn once from the set of ranking algorithms. Therefore, each evaluation measure uses the same set of rankings to remove potential biases from interactions between rankings and evaluation measures.

The results in Fig. 12 show that most evaluation measures perform similarly independent of how many ranking algorithms are compared. The evaluation measure that varies the most is the Min measure while all other measures form relatively dense clusters. The clusters are labelled to indicate the test data set and database used, where the first label indicates the test data set and the latter indicates the database.

The formation of dense clusters shows that the evaluation measures are relatively stable when the number of ranking algorithms are varied. This is important since often in experimental settings only a small number of metrics are analysed and compared. It also indicates that the framework is stable. The results only vary between test data sets and publication databases and remain relatively stable when different rankings are used.

The measures AP and PR have very similar results with PR yielding slightly better results. For space reasons we only plot the PR results. For measures that define cut-offs we only show the results of the variant for which the best results are obtained.

The measures that do not perform well are Min, ROC@Avg, Recall@2000 and Max. The performance of nDCG is substantially improved when the average rank of the relevant entities is used as a cut-off threshold. All other measures perform reasonably well.

7. Threats to validity and future work

7.1. *Internal validity*

The entities in the ACM fellows test data set are unique. This is not true for the LCA data set since some researchers have won more than one lifetime contribution award. We therefore only use entities once in the LCA data set even if the corresponding researchers won more than one award. However, it should be noted that these two data sets are not completely disjoint. About 20% of all ACM fellows have also won at least one LCA award. On the ACM and MAG databases the overlap is respectively 198 and 189 entities. For the BPA and the HI data sets the overlap is very small where 15 papers overlap on the ACM database and 13 on the MAG database. This means that about 2.5% of best papers have also won a high-impact award.

The aim of this paper was to answer the question of which evaluation measures are best suited to evaluate rankings using test data in typical experimental settings. We argue that the overlap of the ACM fellows and LCA authors data sets is not a concern since each data set represents one instance of a typical test data set that a researcher would use in experiments. Furthermore, we show all results separately for the ACM fellows and LCA authors.

7.2. *External validity*

Generalisation of the results is a concern and requires further investigation. The results showed that the evaluation measures are relatively volatile to changes in the number of queries used in the framework. The number of queries ultimately depends on the size of the test data that is available. The larger the test data the more queries can be constructed with more relevant entities per query. For rigorous experimentation we suggest that this parameter is analysed each time.

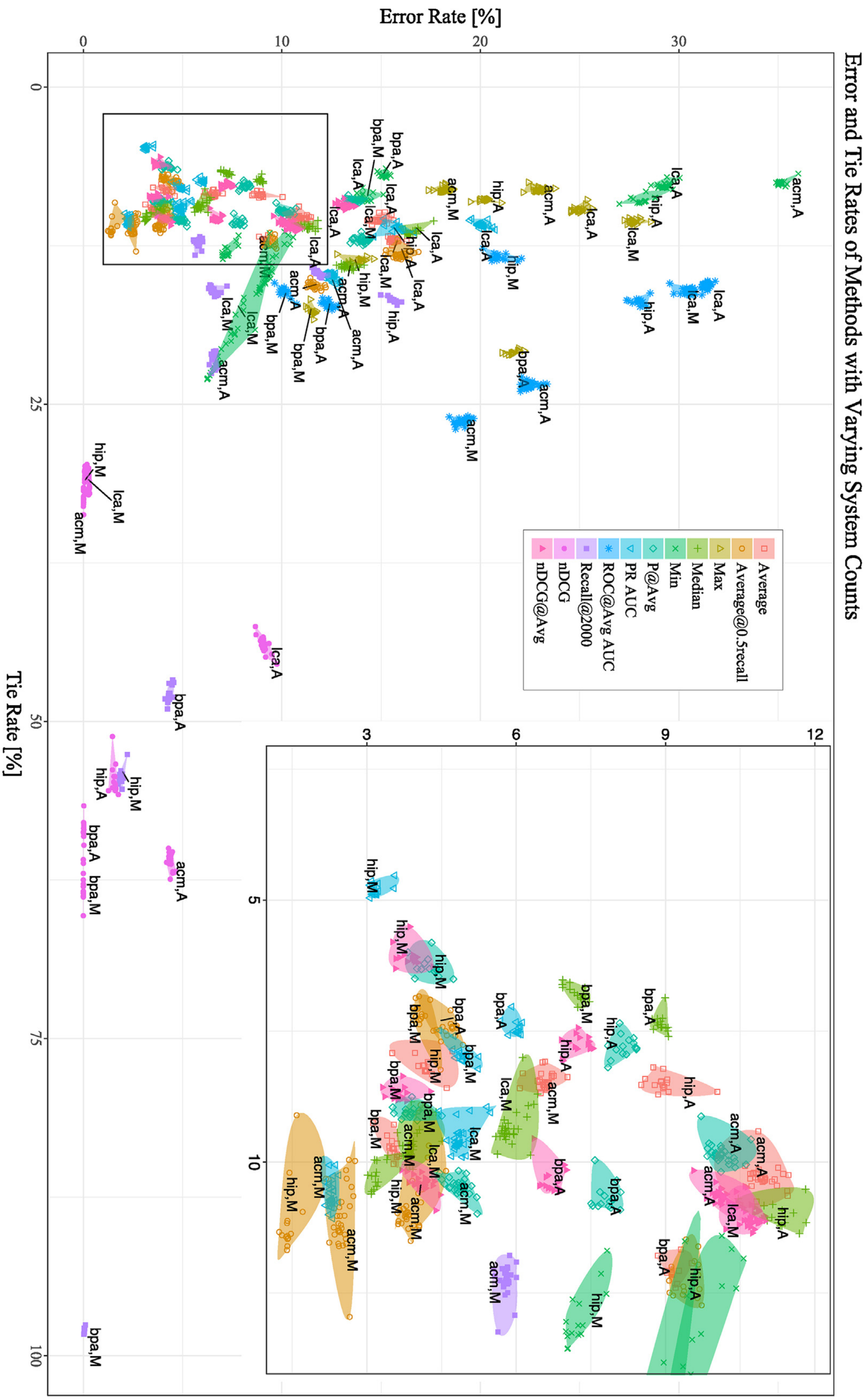


Fig. 12. The average error and tie rates for the evaluation measures with varying number of rankings that are used for the computations. For paper and author test sets the rankings are randomly sampled from 38 and 64 different ranking algorithms. The first label of each cluster indicates the test data set while the latter indicates whether the ACM (A) of MAG (M) database is used.

All rank distributions of the test data entities are positive skew and long-tailed to the right (as depicted in Fig. 2). This is true for all ranking algorithms, as well as the rank distributions of permille rankings, since they are ultimately based on citation counts. When test data with notably different rank distributions are used, some results will likely change. For example, if a left-skewed distribution is used, a cut-off at the average rank might not be a good choice since more relevant entities will be ignored.

The rankings of the test entities are also very sparse since the size of the test data sets are orders of magnitudes smaller than the number of entities in the databases. The number of test entities is always smaller than 1000 while the author and paper entities in the ACM (MAG) database are respectively 625 566 (32 million) and 1 038 063 (18 million). Therefore, the sparseness of the rankings differ substantially between the different databases and we showed that the results generalise as such. However, we did not investigate which evaluations measures perform well on dense rankings.

It should also be noted that measures for which a cut-off threshold is based on the relevant entities, such as Average@Avg, the cut-off thresholds are computed separately for each query. In other words, the cut-off thresholds of queries during the same experiment vary depending on the relevant entities in the corresponding queries. An alternative approach is to compute the cut-off thresholds first using all relevant entities and then computing the measure's scores for each query using the same cut-off. This approach basically defines a static cut-off such as Average@2000 except that the cut-off is defined by the ranks of the relevant entities. We did not investigate whether this alternative approach has a significant impact on the results.

Lastly it should be mentioned that the ASL rate is calculated based on the aforementioned varying cut-offs while the overall differentiation rate uses a single cut-off which is the average rank of all relevant entities. As a consequence, the ASL rate is not always directly proportional to the overall differentiation rate. In general, the ASL rate should be interpreted as the general sensitivity of a measure and used to compare measures. The differentiation rate should be interpreted as one instance of achieved differentiation performance in an application setting.

7.3. Future work

The proposed framework is adapted from methods previously used in typical text-based information retrieval environments. In those experiments the researchers used many different types of query sets which allowed them to evaluate measures such as the mean average precision (MAP). MAP computes the average score of multiple average precisions over a set of queries. To evaluate this type of evaluation measure multiple query sets of different query set types are required.

It is possible to construct artificial query set types using the data presented in this paper. For example, eight query set types could be constructed through the combination of test data set and publication database. Using this approach the stability of MAP could be directly evaluated and compared against, for example, the mean nDCG measure over multiple queries.

8. Conclusion

In this paper we proposed a framework for evaluating evaluation measures. It can be used to identify the best evaluation measure when ranking algorithms are evaluated with test data. Furthermore, it can be used to estimate the performance difference required between two or more ranking algorithms to judge one algorithm better than another. We found that evaluating sparse rankings is a difficult task and that there is no straight-forward answer to which evaluation measure should be used.

We pointed out that the conversion of metrics' scores to ranks is not trivial and argue that fractional ranking is the fairest approach when comparing rankings of different algorithms. Furthermore, we proposed a new formulation of how common evaluation measures can be adapted to score percentile rankings. Using permille rankings has the benefit of normalising rankings between databases with different sizes. Other than that, we found no clear advantage to using permille rankings since the stability and discriminative power of most evaluation measures is not significantly improved. The exception is the nDCG measure without a cut-off which only performs well when permille rankings are used.

In general, when cut-off values have to be specified, we suggest that they are not defined by static values or by the number of relevant entities in test data sets. Instead they should be based on the ranks of the relevant entities. We found that specifying the cut-off threshold at the average rank of the relevant entities or the rank at which 50% recall is achieved, performs the best.

Even though no best evaluation measure is identified, we made a few interesting observations. In general, the discriminative power of nDCG is very low but can be significantly improved by choosing appropriate cut-off thresholds. The sensitivity of Precision@n and Recall@n is highly dependent on the cut-off threshold. The attentive reader might have noticed that R-precision was never mentioned in the results. The R-precision measure is identical to Precision with a cut-off threshold defined by the size of the test data set. We found that this cut-off threshold is too small to produce good results not only for Precision but for all measures that require a cut-off. However, we found that Precision with a cut-off threshold defined by the average rank of the relevant entities performs well.

We showed that most of the common evaluation measures have high stability and discriminative power. The measures with the most stable results and consistently high discriminative power are Precision and nDCG with a cut-off at the average rank, as well as the Average rank measure.

The score differences in average and median ranks required to confidently differentiate rankings is surprisingly high. For example, using BPA papers as test data on the ACM database, a minimum difference of over 9000 in the average rank is required to differentiate between two algorithms with 95% confidence. Yet we found that, most of the time, this required difference is smaller than the actual differences produced by our sample of ranking algorithms.

Author contributions

Marcel Dunaiski: Conceived and designed the analysis; Collected the data; Contributed data or analysis tools; Performed the analysis; Wrote the paper.

Willem Visser: Other contribution supervision and reviewing.

Jaco Geldenhuys: Other contribution Supervision, reviewing and proof-reading

References

ACM, Inc. (2014). ACM Digital Library.

Altman, A., & Tennenholtz, M. (2010). An axiomatic approach to personalized ranking systems. *Journal of the ACM*, *57*(4), 1–35.

Bouyssou, D., & Marchant, T. (2016). Ranking authors using fractional counting of citations: An axiomatic approach. *Journal of Informetrics*, *10*(1), 183–199.

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval*, 33–40.

Cleverdon, C. (1967). The Cranfield tests on index language devices. *ASLIB proceedings, vol. 19*, 173–194.

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).

Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, *10*(2), 392–407.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). Author evaluation at scale. *Journal of Informetrics* (submitted for publication)

Fiala, D., & Tutoky, G. (2017). PageRank-based prediction of award-winning researchers and the impact of citations. *Journal of Informetrics*, *11*(4), 1044–1068.

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), 334–348.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks. *Journal of Informetrics*, *6*(3), 370–388.

Gao, C., Wang, Z., Li, X., Zhang, Z., & Zeng, W. (2016). PR-Index: Using the h-index and PageRank for determining true impact. *PLOS ONE*, *11*(9), e0161755.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, *20*(4), 422–446.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*(4), 1207–1223.

Microsoft. (2017). *Microsoft academic graph*. (accessed 15.08.17). https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

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

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

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 525–532.

Shi, H., Tan, Y., Zhu, X., & Wu, S. (2013). Measuring stability and discrimination power of metrics in information retrieval evaluation. *International conference on intelligent data engineering and automated learning*, 8–15.

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

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators. *Scientometrics*, *108*(1), 337–347.

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02* (pp. 316–323).

Voorhees, E. M. (2002). The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*. pp. 355–370. Berlin Heidelberg: Springer.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators. *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379.

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.

Journal of Informetrics 13 (2019) 270–290

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

On the interplay between normalisation, bias, and performance of paper impact metrics

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[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 14 October 2018
Received in revised form 9 January 2019
Accepted 9 January 2019

Keywords:
Impact indicators
Ranking evaluation
Field normalisation
Field bias
Test data

ABSTRACT

We evaluate article-level metrics along two dimensions. Firstly, we analyse metrics' ranking bias in terms of fields and time. Secondly, we evaluate their performance based on test data that consists of (1) papers that have won high-impact awards and (2) papers that have won prizes for outstanding quality. We consider different citation impact indicators and indirect ranking algorithms in combination with various normalisation approaches (mean-based, percentile-based, co-citation-based, and post hoc rescaling). We execute all experiments on two publication databases which use different field categorisation schemes (author-chosen concept categories and categories based on papers' semantic information).

In terms of bias, we find that citation counts are always less time biased but always more field biased compared to PageRank. Furthermore, rescaling paper scores by a constant number of similarly aged papers reduces time bias more effectively compared to normalising by calendar years. We also find that percentile citation scores are less field and time biased than mean-normalised citation counts.

In terms of performance, we find that time-normalised metrics identify high-impact papers better shortly after their publication compared to their non-normalised variants. However, after 7 to 10 years, the non-normalised metrics perform better. A similar trend exists for the set of high-quality papers where these performance cross-over points occur after 5 to 10 years.

Lastly, we also find that personalising PageRank with papers' citation counts reduces time bias but increases field bias. Similarly, using papers' associated journal impact factors to personalise PageRank increases its field bias. In terms of performance, PageRank should always be personalised with papers' citation counts and time-rescaled for citation windows smaller than 7 to 10 years.

1. Introduction

Citation metrics constitute a key tool in scientometrics and play an increasingly important role in research evaluation (Bornmann, 2017). To enable fair evaluations, it is a de facto requirement that field-normalised metrics are used (Waltman, 2016). This stems from the observation that not all citations are equal. Citation patterns vary between academic disciplines (Lundberg, 2007) and heterogeneities are even found in narrow subfields within the same discipline (van Leeuwen & Calero

[*] Corresponding author at: Media Lab, Stellenbosch University, 7602 Matieland, South Africa.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

https://doi.org/10.1016/j.joi.2019.01.003
1751-1577/© 2019 Elsevier Ltd. All rights reserved.

Medina, 2012). Furthermore, the scientific corpus steadily grows at an increasing rate but rates fluctuate between different fields, which also contributes to varying citation distributions over time (Bornmann & Mutz, 2015). Therefore, citation metrics should always be evaluated for their fairness when their application is the evaluation of units across fields or time.

Another important aspect is to validate that metrics fulfil their intended purpose. According to Bornmann and Marx (2018), situations should be created or found in empirical research in which a metric can fail to achieve its purpose. A metric should only be regarded as provisionally valid if these situations could not be found or realised. For example, metrics that are intended to rate the quality of papers should be assessed by correlating them with peer assessments (Bornmann & Marx, 2015). However, collecting direct peer-assessed test data is time consuming and expensive. We therefore use a proxy for this assessment which comprises test data based on other ground truth provided by domain experts (Dunaiski & Visser, 2012; Dunaiski, Visser, & Geldenhuys, 2016; Mariani, Medo, & Zhang, 2016; Sidiropoulos & Manolopoulos, 2005). Specifically, we use selected papers as test data that have won prizes for their outstanding quality and papers that have won awards for their continued high impact in their fields. In this paper we follow the appeal by Bornmann and Marx (2018) for continued scrutiny of current proposals and evaluate article-level metrics along two dimensions: (1) their fairness to rank papers across fields and time, and (2) their performance in identifying the relevant entities in our test data sets.

Since no gold standard for normalised metrics exists (Bornmann & Marx, 2018), we evaluate a variety of different approaches. We evaluate mean-based (Radicchi, Fortunato, & Castellano, 2008) and percentile-based (Bornmann, Leydesdorff, & Mutz, 2013) metrics based on citation counts that normalise scores over fields and publication years. We also evaluate the relative citation ratio (RCR) metric that normalises a paper's score by its co-citation network and age (Hutchins, Yuan, Anderson, & Santangelo, 2016). Furthermore, we evaluate citation counts and PageRank scores that are rescaled across fields and time intervals (Vaccario, Medo, Wider, & Mariani, 2017).

We conduct all experiments on two publication databases. The first database is the ACM Digital Library (ACM Inc., 2014) and use its Computing Classification System (CCS) to analyse the metrics' ranking bias on subfields in the computer science discipline. The CCS consists of a library-like, hierarchical structure of concepts where papers are assigned to one or more concepts by their authors. The second database is the Microsoft Academic Graph (MAG) database (Microsoft, 2017). It is multi-disciplinary and papers are assigned to fields in a hierarchical structure based on keywords extracted from their texts. We use the top-level fields as categories which roughly capture the scientific disciplines such as 'Mathematics' or 'Medicine'.

With this paper we make the following contributions:

- We extend the test data driven performance validation of Mariani et al. (2016) to additional paper metrics using substantially larger test data sets and two different publication databases.
- We analyse the bias and performance of paper metrics with test data that consists of papers from different fields. This allows us to evaluate the normalised metrics across fields.
- Our evaluation also uses two different types of field categorisation schemes. The first is a categorisation where the authors chose their papers' categories (ACM database). The second is based on semantic information contained within papers' titles and abstracts (MAG database).
- Lastly, we also analyse how well the various metrics identify high-impact papers shortly after their publication date while considering the intrinsic ranking characteristics of metrics.

In this paper, we first provide the reader with background information about normalisation factors, bias evaluation, and the paper metrics we use (Section 2). In Section 3 we describe the methodology of evaluating the metrics along the bias and performance dimensions. We present the results in Section 4, followed by a discussion on the results in Section 5.

2. Background information

2.1. Normalising factors

There are various properties of publication data that should be utilised as normalisation factors to ideally correct for the imbalance of citation opportunity. On the paper level, arguably the most important factor to consider are academic fields. Different fields have varying citation and publication cultures which result in different mean citation counts between fields (Lundberg, 2007; Radicchi et al., 2008).

To overcome this bias, mean-based metrics have been suggested that normalise papers' citation scores by the average citation score in a field (Radicchi et al., 2008). However, citation distributions of papers are inherently skewed, with many papers that only obtain a few citations or none at all, and long-tailed (only a few papers receive a large number of citations). Comparing papers in the long tail is not trivial since it is difficult to say how many citations constitute a large enough number to signify a meaningful difference (Ioannidis, Boyack, & Wouters, 2016). Therefore, percentile-based metrics have been proposed where the citation score of a paper is rated in terms of its percentile in the citation distribution of the field to which it belongs (Bornmann et al., 2013; Leydesdorff, Bornmann, Mutz, & Opthof, 2011).

The difficulty in normalising for fields is how to assign papers to fields. In the past, fields have been categorised on the basis of journals or library categories. The problem is that fields and especially broad disciplines are not isolated. Generally, within-field citations are denser than between-field citations, however, between-field citations are becoming more common nowadays (Silva, Rodrigues, Oliveira, da, & Costa, 2013). For example, it has been shown that by considering only the most aggregate level (i.e., disciplines), heterogeneities in the subfields' citation patterns might be disregarded (van Leeuwen & Calero Medina, 2012). This is more problematic for indirect metrics (PageRank) compared to citation counts (Waltman, Yan, & van Eck, 2011). An agreement about the optimal classification scheme has yet to be reached (Adams, Gurney, & Jackson, 2008; Colliander & Ahlgren, 2011; Zitt, Ramanana-Rahary, & Bassecoulard, 2005).

Alternatives to defining fields a priori exist. Fields may be defined dynamically based on the citation network structure or the semantic relatedness of papers' contents. Hutchins et al. (2016), for example, use a paper's co-citation network to define its field. The rationale is that if papers that are cited together by another paper, they belong to the same topic since they were relevant in producing the new paper. However, this does not hold for all citations. Janssens, Goodman, Powell, and Gwinn (2017) found that most of the co-cited papers that are in a paper's co-citation network due to only a few co-citations do not belong to the same topic. Furthermore, some work is inherently of such a nature that it attracts citations from multiple, generally unrelated fields, such as statistical methods (Silva et al., 2013).

Normalising over ill-defined fields may also lead to undesirable situations. For example, when a paper's field is defined by its co-citations and it receives new citations from a remote field it may indicate an increase in importance. However, as Waltman (2015) points out, if the remote field has a high citation density, normalisation may lead to a decrease rather than an increase in the paper's score with the newly acquired citations. Similarly, defining categories by individual journals (Pudovkin & Garfield, 2009) is also problematic and normalisation over such categories has to be well-justified. For example, the top paper from a prestigious journal should probably not be scored the same as the top paper from an obscure (generally under-cited) journal.

In this paper, we use two different categorisation schemes. On the ACM database, we use the ACM classification system (ACM Inc., 2017) which consists of library-like categories that are fine-grained and author-chosen. We also use the MAG field categorisation scheme which is based on the semantic information contained in papers' keywords, titles, and abstracts (Microsoft, 2017). In our analyses, we also use various metrics that define fields differently for their normalisation steps. For example, the RCR metric uses papers' co-citations (Hutchins et al., 2016), PRI uses journals (Pudovkin & Garfield, 2009), while others use the above-mentioned field categories.

The age of publications should also be considered for normalisation (Ioannidis et al., 2016). Older papers have had more time to accrue citations and therefore papers should only be compared directly to papers that were published around the same time. However, since the scientific corpus is growing at an increasing rate (Bornmann & Mutz, 2015), older influential papers had a lower citation potential shortly after their publication compared to younger influential papers. Furthermore, Ioannidis et al. (2016) argue that for any field, there is a fluctuation of productivity from year to year. For instance, a paper in the top 10% of cited papers in a year of major progress may be more important than a paper in the top 1% of a year when the field stagnated.

Lastly, time normalisation based on calendar years produces noisy results for the most recently published papers (Ioannidis et al., 2016). Parolo et al. (2015) show that for all ages, the number of papers is a better indicator to capture the role of time in academic citation networks than actual time. Papers are labelled in order of decreasing age and separately for each field. Therefore, the oldest paper receives a new label 0 while the newest paper in a field receives the label n (Newman, 2009). Accordingly, Mariani et al. (2016) propose to rescale a paper's score $S(p_i)$ by calculating the mean score μ_i and standard deviation σ_i of the closest papers in the same field as paper p_i. The closest papers to p_i are the papers with the labels $j \in \left[\max\{i - \Delta/2, 0\}, \min\{i + \Delta/2, n\}\right]$, where Δ is the size of the considered time interval.

Therefore, the normalised score of a paper is the following:

$$S_{norm}(p_i) = \frac{S(p_i) - \mu_i}{\sigma_i} \tag{1}$$

It should be noted that in terms of time bias, simply computing the ratio (i.e., $S(p_i)/\mu_i$) for rescaling performs worse since it strongly depends on the age of papers (Mariani et al., 2016). We therefore use Eq. (1) to rescale metrics that do not incorporate any normalisation. For example, we indicate time-rescaled citation counts as 'Citations[Δ]'. Similarly, we indicate time- and field-rescaled citation counts as 'Citations[F,Δ]'.

2.2. Quantifying ranking bias

According to Radicchi and Castellano (2012) the fairness of a metric is "directly quantifiable by looking at the ability of the [metric] to suppress any potential citation bias related to the classification of papers in disciplines or topics of research". The idea of the fairness test for metrics is to measure the deviation of the field distribution of the top p percent of papers to the field distribution of the overall sample of papers. For example, given a sample of 80 computer science and 20 mathematics papers, a fair metric would score the papers in such a way that the top 10% of papers comprise 8 computer science and 2 mathematics papers. However, it is important to note that this fairness test may be biased itself if the category scheme used to normalise a metric is identical to the one used for measuring its fairness (Sirtes, 2012).

Vaccario et al. (2017) propose a method to quantify ranking bias based on the fairness test but which also allows to compute overall confidence intervals to judge whether ranking biases are statistically significant. In addition, using their approach one may compute a comparable per-field impact on the overall bias of a ranking. It uses the Mahalanobis distance (d_M) (Mahalanobis, 1936) to quantify the deviation between two distributions and is based on the assumption that a ranking is unbiased if its properties are consistent with that of an unbiased sampling process (Vaccario et al., 2017). Therefore, a percentage p from all papers N are randomly drawn without replacement. Based on this sample, the frequency that a field is observed is recorded in a vector $\boldsymbol{k}$.

Let n be the number of papers constituting p% of all papers. Then the multivariate hypergeometric distribution gives the probability of observing such a vector $\boldsymbol{k}$

$$P(\boldsymbol{k}) = \frac{\prod_{f \in F} \binom{K_f}{k_f}}{\binom{N}{n}} \tag{2}$$

where F is the set of fields and K_f is the total number of papers in field f. An unbiased ranking would yield the expected number of papers for a field f as $\mu_f = n \cdot K_f / N$. Let $k_f^{(m)}$ be the number of top p% of papers in field f from an actual ranking metric m. To quantify the deviation of the observed vector $\boldsymbol{k}^{(m)}$ from the expected vector $\boldsymbol{\mu}$, Vaccario et al. (2017) propose to simulate n_{sim} number of unbiased selection processes to obtain a set of ranking vectors distributed according to Eq. (2) around the vector of expected values $\boldsymbol{\mu}$. For each vector in this distribution they compute the Mahalanobis distance to $\boldsymbol{\mu}$ to obtain a distribution of Mahalanobis distances from which one can compute confidence intervals.

2.3. Paper ranking metrics

Although using raw citation counts for papers has widely accepted drawbacks for quantitative evaluations, we use them as a baseline metric for comparisons in this paper. Radicchi et al. (2008) show that the probability of papers receiving a citation has a large variation between different disciplines. They use the average citation counts of papers published in the same field and the same year to normalise paper citation counts. Let Citations(p_i) be the citation count of paper p_i. Then the *relative citation count* metric is defined as follows:

$$\text{Citations}_{rel}(p_i) = \frac{\text{Citations}(p_i)}{\mu_f} \tag{3}$$

where μ_f is the mean citation count of papers in field f that were published the same year as p_i. Radicchi et al. (2008) claim that by mean-rescaling citation counts, all field and time categories adhere to the same universal citation distribution. However, their findings have been questioned since they use a small number of sample fields (Albarrán, Crespo, Ortu no, & Ruiz-Castillo, 2011) and exclude un-cited papers (Waltman, van Eck, & van Raan, 2012). Furthermore, Waltman et al. (2012) show that, when including un-cited papers, fields with low average citation counts per paper do not conform to a universal citation distribution. Moreover, Albarrán et al. (2011) show that for the upper and lower tails of citation distributions this universality partially breaks down.

For comparisons we use two percentile approaches. The first is the *percentile rank index* (PRI) proposed by Pudovkin and Garfield (2009) which assigns percentile values to papers based on citation counts per journal+year categories. Let n be the number of papers published in a journal in a certain year and let $R(p_i)$ be the fractional rank of paper p_i. Then PRI is computed as follows:

$$\text{PRI}(p_i) = 100 \cdot \frac{n - R(p_i) + 1}{n}$$

The second percentile approach uses the exponential distribution formula proposed by Gringorten (1963) and recommended by (Bornmann et al., 2013). Instead of using journals as categories, let n be the number of papers in a field+year category. Then this percentile approach is defined as follows:

$$\text{Percentile}(p_i) = 100 \cdot \frac{(n - R(p_i) - 0.44)}{(n + 0.12)}$$

and 0 for un-cited papers.

The *relative citation ratio* (RCR) is a metric that uses the co-citation network of papers to normalise papers' citation counts (Hutchins et al., 2016). The co-cited papers of a paper p_i is the set of papers that are cited together with paper p_i. In other words, all papers that are found together with paper p_i on reference lists constitute p_i's co-citation network. The rationale is that co-cited papers are topically similar since they are cited together. The *actual citation rate* is defined as ACR(p_i) = Citations(p_i)/*age*(p_i), where *age*(p_i) is the number of years since the publication of p_i.

The normalisation factor for a paper is its *field citation rate* (FCR) which is based on the journals in which the papers' co-cited papers are published. Therefore, a journal citation rate (JCR) is computed for every year, which is the number of citations accrued by the journal during a two-year citation window through papers published in a certain year. The rational is that the JCR values for journals are stable over time. The FCR(p_i) is then the average JCR for all papers in p_i's co-citation network. A JCR is counted multiple times if more than one paper from the same journal and year occurs in p_i's co-citation network. Thus, the RCR for paper p_i is

$$\mathrm{RCR}(p_i) = \frac{\mathrm{ACR}(p_i)}{\mathrm{FCR}(p_i)} \tag{4}$$

Instead of only considering direct citations, indirect metrics also consider the indirect impact of papers through reference chains. Most indirect metrics are recursively defined and take the entire structure of citation networks into account. The idea of recursively defining impact metrics originates from Pinski and Narin (1976). They applied it on academic citation networks to compute importance values for journals and to address the limitation that all citations are valued the same. The rationale of applying indirect metrics to citation networks is that citations from influential papers should count more than citations from unimportant papers. This idea was made popular with the introduction of *PageRank* (Brin & Page, 1998). Intuitively, PageRank simulates a process in which random walkers are placed on papers and follow citations to other papers. This continues until they are teleported to new random papers controlled by a teleportation probability $(1-\alpha)$, where α is a parameter of PageRank called the 'damping factor'. If the random walkers reach papers without outgoing references, they restart their searches with new random papers.

Two aspects of this process are important to consider when applying PageRank on paper citation networks. Firstly, the damping factor has to be chosen carefully since it controls the path lengths of the random walkers (Chen, Xie, Maslov, & Redner, 2007). The rankings produced by PageRank are more biased towards older nodes the longer the paths of the random walkers (i.e., $\alpha \to 1$). Secondly, a personalisation vector $\boldsymbol{r}$ controls the initial placement of the random walkers and their restarts. It can be initialised, for example, to skew the probabilities towards papers in a certain field or papers published during a certain time period.

Formally, let A be the adjacency matrix of the citation network containing n papers, where A_{ij} is 1 if paper i cites paper j and 0 otherwise. In order for the PageRank algorithm to converge, A has to be a left stochastic matrix and therefore has to be normalised such that each column's sum is 1. Furthermore, let $\boldsymbol{d}$ be a vector with values $d_i = 1$ if the paper i is a dangling paper (no outgoing references) in the network and 0 otherwise.

The PageRank scores for papers are contained in the vector $\boldsymbol{x}$ in the following equation:

$$\boldsymbol{x} = \underbrace{\frac{(1-\alpha)}{n}\cdot\boldsymbol{r}}_{\text{Random Restarts}} + \alpha\cdot\left(A^T + \underbrace{\frac{1}{n}\cdot\boldsymbol{r}\cdot\boldsymbol{d}^T}_{\text{Dangling Nodes}}\right)\cdot\boldsymbol{x} \tag{5}$$

It should be noted that the PageRank algorithm defined here adds n edges from each dangling node to all other nodes in the graph. The weights associated with these added edges are distributed according to the values in $\boldsymbol{r}$. This is modelled by the 'Dangling Nodes' term in Eq. (5), while the 'Random Restarts' term models the distributed placement of random walkers when they restart. Unless specifically stated otherwise, we set PageRank's damping factor α to 0.5.

The last metric we consider also takes the impact of citing papers into consideration (Giuffrida, Abramo, & D'Angelo, 2018). However, differently to PageRank, it only considers two citation levels and transfers less of the citing paper's impact to the cited paper, by implementing the restriction that the gain through citations from high-impact papers should not be more than 1 (i.e., two citations cannot count less than one). This metric also normalises scores over fields and publication years. For the lack of an existing name we refer to this metric as the *Abramo* method after the name of the paper's corresponding author.

3. Methodology

3.1. Publication databases

We use two publication databases for the experiments described in this paper. The first is a 2015 version of the ACM Digital Library (ACM Inc., 2014). It contains papers up to March 2015 that are published in periodicals and proceedings from the field of computer science. The ACM uses a categorisation scheme which is called the Computing Classification System (CCS) where each paper is associated with one or more concepts that are organised in a poly-hierarchical structure (ACM Inc., 2017). Each concept belongs to one or more parent concepts. Of all papers in the database, 703 802 have at least one associated concept (with an average of 4.49 concepts per paper). Since each concept is part of the hierarchical structure it can be collapsed such that each concept is associated with at least one top-level concept.

There are 13 top-level concepts in the CCS which we use as fields (see Table B.14 in Appendix B). We collapse the CSS concept structure such that papers are only associated with corresponding top-level concepts. A paper's field is then the top-level concept with which the paper is associated most frequently. This decision is based on the assumption that if a paper is associated with a top-level concept multiple times, it is closely related to that top-level concept and can therefore be categorised as such. For papers with equal frequencies of top-level concepts, we categorise them into each of the most occurring top-level concepts. We found that of all papers with concepts, 77% have a single most occurring top-level concept. All papers without concepts we classify into a separate field named 'none'. This decision is based on the fact that too many

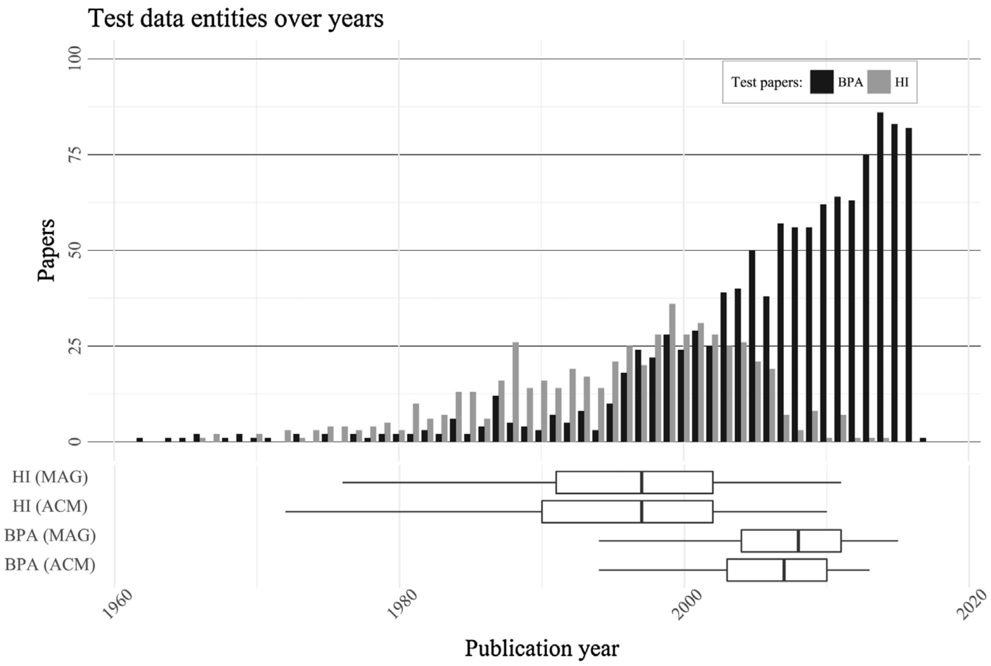


Fig. 1. The top plot shows all BPA and HI test papers over publication years. The bottom box plots summarise the publication year distributions of the two test data sets per database. Papers receive high-impact awards typically 10–25 years after publication, while best paper awards are handed out the year of publication.

citations would be removed if all papers without concepts were removed. After removing papers that are not associated with a journal, conference series, or publication date, the ACM data set comprises 1 737 687 papers.

For the Microsoft Academic Graph (MAG) database, we follow the categorisation method similar to the one used by Vaccario et al. (2017). Hence, it includes all paper types (paper, review, books, etc.) and publication venues (journals, conference, etc.). We remove papers that are not associated with a journal, a conference, or a complete date (yyyy/mm/dd). In addition, for fair comparisons between metrics, we only consider papers that have at least one reference or at least one citation. For papers that do not receive scores by a metric (i.e., papers with no citations), we assign the value of 0. The MAG database includes a topic classification scheme with four hierarchical levels. However, we only use papers that are associated directly to one of the 19 top-level fields which can be interpreted as broad academic disciplines (see Table B.15 in Appendix B for the MAG field sizes), since lower-level fields in MAG are too noisy for citation analyses (Hug, Ochsner, & Brändle, 2017). The final MAG database comprises 13 829 901 papers.

3.2. Test data sets

We use two test data sets which we manually collected from various websites to evaluate the performance of the paper ranking metrics. The first comprises 1119 papers that won best paper awards (BPA) at 36 different conferences or journals. The exact selection process of best paper awards differs between venues, however, the ratings are usually based on papers' intrinsic quality judged by reviewers and final decisions are made by editors or conference committees. Best paper awards are decided before publication and therefore no knowledge about future impact (citations) is known.

The other test data set consists of 563 papers that have won a high-impact (HI) award from one of 30 different conferences or organisations. These awards are handed out post-publication, usually 10–25 years after their initial publication, by selection committees comprising reviewers that can be assumed to be experts in the corresponding fields. Papers are typically evaluated on their continued impact in their field in terms of research, methodology, or application. HI papers are therefore expected to have above average citation counts.

The top plot in Fig. 1 shows how the BPA and HI papers are distributed over the publication years. The HI papers are on average older and only very few are published after 2010, while the BPA are published more frequently in recent years. We matched the papers in the test data sets to their corresponding entities in the ACM and MAG databases by matching their titles or DOIs and validating whether the publications years correspond. After we remove test entities that do not fall into the subset of the databases used (see Section 3.1), we obtain the following test entity counts:

- **BPA papers**: 516 unique papers that won a best paper award are associated with the ACM database and 505 are associated with the MAG database.
- **HI papers**: 401 and 354 high-impact papers are associated with the ACM and MAG databases, respectively.

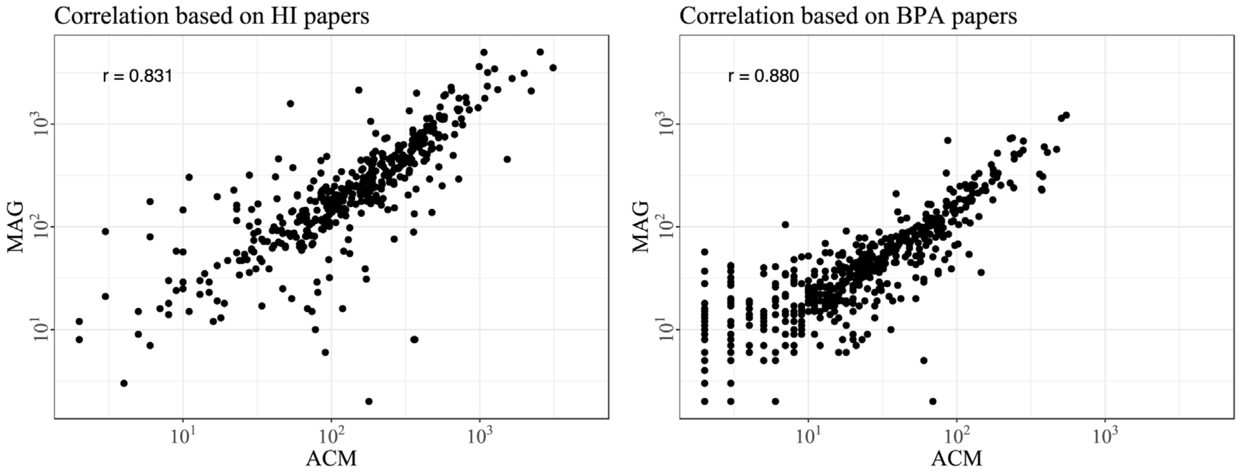


Fig. 2. Pearson correlations of the HI (left) and BPA (right) papers' citation counts between the ACM and MAG databases.

On the MAG database the sizes of the usable test data sets are substantially reduced since many papers are not associated with top-level fields in the database. The box plots in the bottom of Fig. 1 summarise the publication year distributions of the two test data sets after matching their entities to the ACM and MAG databases. They highlight the difference in the publication year distributions between the two test data sets.

We assume that the entities in the test data exhibit some property that is not exclusively based on citations. For the HI papers the assumption is that they have had a long-lasting and influential impact on future papers. Therefore, we expect HI papers to have above average citation rates but also to have a latent property that is not encoded through pure citations. For the BPA papers the underlying property is that they are high quality but might not have high impact. There are many other factors that influence the decisions of awarding best paper prizes that are not measurable through citations. Therefore, this underlying property is further detached from citation counts compared to the HI papers. We use these two distinct test data sets to investigate whether certain metrics better identify the test entities and consequently their underlying property.

We use the papers that are matched to both databases in Fig. 2 to show the citation count correlations between the ACM and MAG databases. The left and right plots show the correlations based on the HI and BPA papers' citation counts, respectively. The citation counts are relatively highly correlated (Pearson coefficient of 0.83 for the HI papers and 0.88 for the BPA papers) despite the fact that the ACM database only contains internal citations where all citations from papers that are not indexed by the ACM are excluded. Lastly, it should be noted that the HI papers do have more citations on average (435.09 and 250.40 on the MAG and ACM databases) compared to the BPA papers (72.60 and 43.67, respectively). In Appendix B we show that the BPA papers exhibit some latent property that differentiates them from an "average" paper which some metrics identify better than others.

3.3. Evaluation

For the bias evaluation of the metrics, we use the methodology by Vaccario et al. (2017) which we briefly described in Section 2.2. For every experiment we simulate 1 000 000 unbiased sampling processes in which $p = 1\%$ of papers are sampled to create the statistical null model to which the metrics are compared. To evaluate time bias, we split all papers into 40 equally sized chunks that form the groups from which papers are selected in the sampling process. Similarly, when we compute the null model for a field+time unbiased ranking, the papers are first grouped into 40 time groups and then separated into corresponding fields. For example, the MAG database has 19 fields and therefore the null model comprises 760 groups. Each paper is considered multiple times in the bias analyses, once for each field it belongs to.

The concepts (ACM) and fields (MAG) overlap. One solution is to compute expected scores for fields where a paper that belongs to fields X and Y is evenly attributed to both fields (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). Since the sum of all paper scores between fields X and Y are not equal, the paper's score has to be attributed to field X and Y proportionally (Smolinsky, 2016). Alternatively, if the score of the paper is evenly attributed to fields X and Y, then the paper has to be attributed to the fields proportionally since the paper counts for fields X and Y are not equal.

We use a different approach called multiplicative counting (Albarrán et al., 2011; Herranz & Ruiz-Castillo, 2012). Instead of dividing a paper or its score into multiple fields, we consider each field independently and assign the whole paper and its score to each field it belongs to. We choose this approach for two reasons. Firstly, the bias computation is based on multiple scores per paper (Vaccario et al., 2017). The rationale is that if a paper is a top paper in field X but only average in field Y, it should be considered a top paper in general in the sampling process. Therefore, in order to fairly evaluate a metric's performance along side its field bias, we keep the fields separate. Secondly, both above-mentioned solutions require the calculations of expected scores for each field and each year under consideration, which is computationally expensive. Furthermore, the rescaling approach in Eq. (1) uses unique paper reference sets for each paper, which makes the above-mentioned approaches

Table 1
The metrics' normalisation strategies and output formats.

Metric	Normalisation	Output
Citations, PageRank	None	Single score
Citations[Δ], PageRank[Δ]	Time ($\Delta = 1000$)	Single score
RCR	Co-citations and year	Single score
PRI	Venue and year	Single score
Abramo, Percentile, Rel. Citations	Field and year	Score per field
Citations[F,Δ], PageRank[F,Δ]	Field and time ($\Delta = 1000$)	Score per field

computationally impractical. The only drawback of this decision is that we cannot compare metrics that normalise across fields directly to metrics that do not. Therefore, we have two scenarios for the performance evaluation of the metrics. First, we evaluate metrics that compute a single score for each paper such as standard citation counts or the RCR metric. Second, we evaluate metrics that normalise across fields.

To fairly compare the various metrics, their output scores have to be converted to ranks. We convert a metric's scores to fractional ranks where papers with tied scores are assigned their average rank (Dunaiski, Geldenhuys, & Visser, 2018). For example, a set of papers with citation counts of {10, 5, 5, 1} are assigned the corresponding ranks of {1, 2.5, 2.5, 4}. In the case of single-score metrics, we simply convert the entire list of scores to a single list of fractional ranks. For the metrics where papers receive a score for each field, we treat the fields independently and compute a list of fractional ranks for each field.

The evaluation of a metric's performance therefore requires us to convert either one or multiple rank distributions of the test papers to a single-value performance score. For single-score metrics we use the average rank (AR) of the test papers as evaluation measure and use the methodology proposed by (Dunaiski et al., 2018) to compute significance values for the differences in performance scores. In the case of multiple rank distributions per metric, we use a weighted mean average rank (MAR$_w$) where the average rank of each field distribution is weighted proportionally to the number of its test entities and averaged over all fields. We give the definitions of these evaluation measures in Appendix A where we also discuss the reasons for choosing these evaluation measures.

We also use the methodology proposed by Mariani et al. (2016) to analyse the performance of the metrics in identifying the test papers while considering the age of papers. In other words, which metrics better identify the papers t years after publication. The rank $r_i(m, t)$ of a test paper i according to metric m is computed t years after publication. Similarly, the best rank of i by all considered metrics (m') is computed $\min_{m'}\{r_i(m', t)\}$. The ratio of these two values for each paper is averaged to yield the *average ranking ratio* $\bar{r}(m, t)$ of metric m for each year t. The parameter t is the upper bound of the considered citation window. Therefore, only citations that originate from papers published up to t years after paper i's publication year are considered. The lower the ratio, the better a metric identifies the test papers. An optimal metric (with an average ranking ratio of 1 for all t) means that it always ranks the test papers higher than every other metric at each citation windows size t (see Eq. (B.1) for the mathematical definition of the average ranking ratio).

4. Results

We evaluate the metrics on their overall ranking performance, bias (field and time), and early identification of high-impact papers. Table 1 briefly summarises the metrics' normalisation strategies and the score outputs they produce.

In Section 4.1 we analyse the bias of all metrics. In Section 4.2 we evaluate the ranking performance of metrics that produce a single score per paper. We compare the RCR and PRI metrics, which incorporate some type of field and time normalisation, to standard citation counts and the PageRank algorithm. Furthermore, we analyse the impact that time normalisation has on the performance of citation counts and PageRank. Lastly, we analyse which metrics best identify the HI papers t years after their publication. In Section 4.3 we show the performance results of metrics that produce multiple scores per paper. We compare citation counts and PageRank, after rescaling them across fields and time intervals, to the other metrics that incorporate field and time normalisations. In Section 4.4 analyse the sensitivity of PageRank's field and time bias on its damping factor and for different personalisation strategies.

4.1. Bias of metrics

The scatter plots in Fig. 3 show the time bias against the field bias of all metrics on the ACM (left) and MAG (right) databases. The vertical and horizontal dashed lines represent the upper bound of the 95% confidence intervals of unbiased rankings with respect to time and fields, respectively. For the ACM (MAG) database these upper bounds are at 7.62 (7.62) for time unbiased rankings and 4.97 (5.62) for field unbiased rankings. Table 2 shows the corresponding bias values for the single-score metrics. Table 3 lists the field bias values and the field+time bias values for the metrics that are field and time normalised.

Considering the time bias on the ACM database in Table 2, the PRI (46.65) and RCR (54.77) metrics have less bias compared to the standard citation counts (97.91) and PageRank (141.32). Time rescaling PageRank and citation counts using $\Delta = 1000$ improves their time bias to 10.75 and 7.39, respectively. The bias value of Citations[Δ] falls below the upper bound of the

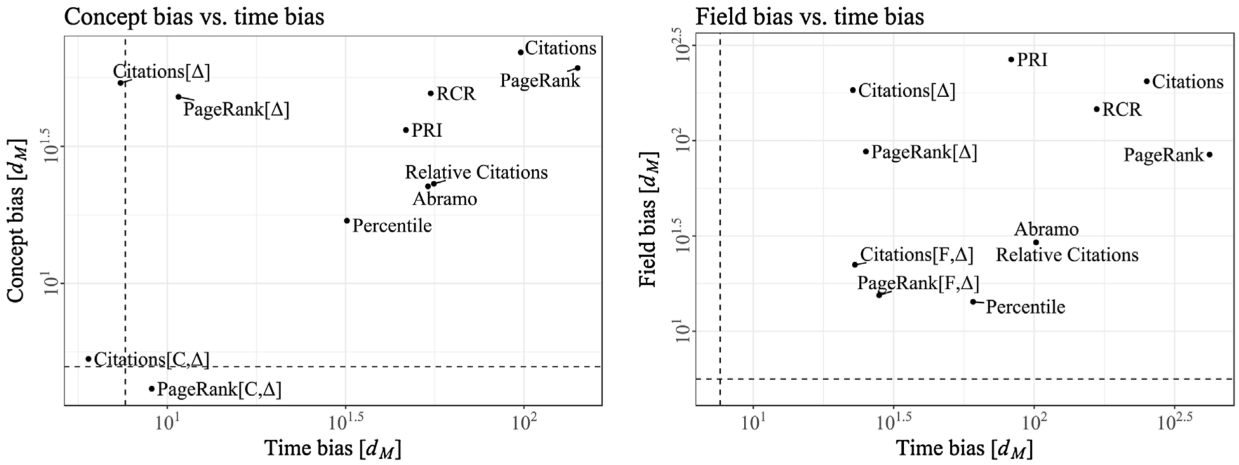


Fig. 3. Time and field bias of the indicators. The left plot shows the concept vs. time bias of the metrics on the ACM database, while the right plot shows their field vs. time bias on the MAG database. The values in brackets indicate the post hoc rescaling used where [Δ] indicates time normalisation (Δ = 1000) while [C,Δ] and [F,Δ] indicates that the metrics are rescaled using time and, respectively, ACM concepts and MAG fields.

Table 2
The time and field (concept) bias of the single-score metrics on the MAG (ACM) database. The last row shows the upper bound of the 95% confidence intervals of unbiased metrics.

Metric	d_M (Time ACM)	d_M (Time MAG)	d_M (Concepts ACM)	d_M (Fields MAG)
Citations	97.91	251.64	69.55	204.91
Citations[Δ]	7.39	22.63	53.80	184.21
PageRank	141.32	420.95	63.49	84.52
PageRank[Δ]	10.75	25.19	47.89	87.68
RCR	54.77	166.96	49.31	146.31
PRI	46.65	82.86	36.24	266.55
UB: 95% CI	7.62	7.62	4.97	5.62

Table 3
The bias values for field+time normalised metrics. We use time splits of T = 40 for the bias computations. The last row shows the upper bound of the 95% confidence interval for unbiased rankings.

Metric	d_M (Concepts ACM)	d_M (Fields MAG)	d_M (Concepts+Time ACM)	d_M (Fields+Time MAG)
Rel. Citations	23.08	29.21	71.60	167.56
Abramo	22.59	29.29	70.22	167.65
Percentile	16.93	14.27	46.90	101.13
Citations[F,Δ]	5.31	22.33	32.72	72.06
PageRank[F,Δ]	4.13	15.43	31.12	51.24
UB: 95% CI	4.97	5.62	25.08	28.94

95% confidence interval (7.62) which means that the rankings produced by Citations[Δ] are unbiased in terms of time. The time bias of the metrics follow the same order on the MAG database. Again, the time-normalised citation counts (22.63) and PageRank (25.19) exhibit the least bias, followed by PRI (82.86) and RCR (166.96).

Considering the concept biases of the metrics on the ACM database, PRI (36.24) and time-rescaled PageRank (47.89) have the least bias, followed by RCR (49.31) and citation counts (53.80). On the MAG database, however, the standard PageRank algorithm is the least biased (84.52), followed by time-rescaled PageRank (87.68) and RCR (146.31). It should be noted that time rescaling citation counts and PageRank also improves their concept bias on the ACM database. On the MAG database, however, the field bias only reduces for citation counts and remains about the same for PageRank.

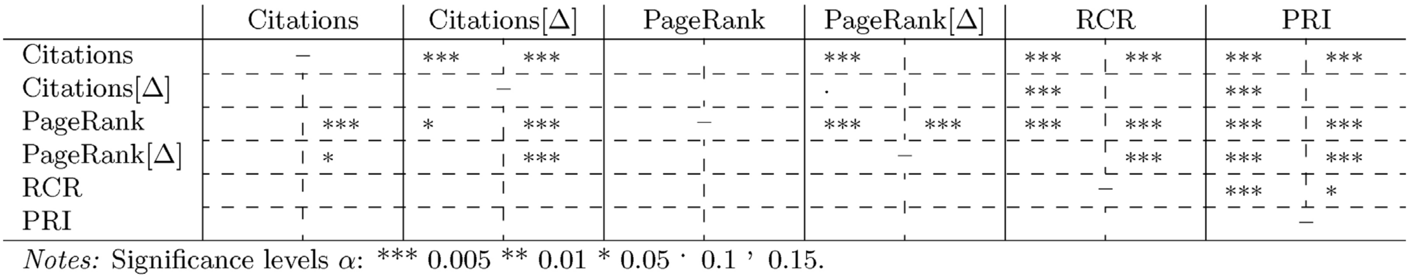
Table 3 lists the bias values of the field+time normalised metrics. Considering the concept bias on the ACM database, PageRank shows the least bias (4.13) and falls below the upper bound of the 95% confidence interval (4.97) of an unbiased ranking. The second least biased ranking is produced by rescaled citation counts (5.31), followed by the percentile approach (16.93), the Abramo method (22.59) and relative citation counts (23.08). The same order of bias values are observed when considering the concept+time biases. However, no metric falls into the 95% confidence interval (25.08) of an unbiased ranking. Compared to the ACM database, the metrics' field biases on the MAG database are slightly different. The percentile approach performs much better with a Mahalanobis distance of 14.27, followed by rescaled PageRank (15.43), and rescaled citation counts (22.33). However, when considering the field+time biases, Citations[F,Δ] and PageRank[F,Δ] are substantially less biased than the percentile approach.

In summary, the most biased metrics are standard citation counts and PageRank, where citation counts are more field biased and PageRank is more time biased. Of the metrics that incorporate field and year normalisations, the Percentile

Table 4
The performance (AR) of the metrics that produce a single score per paper.

Metric	ACM		MAG	
	AR (HI)	AR (BPA)	AR (HI)	AR (BPA)
Citations	9933	96 814	581 642	3 289 763
Citations[Δ]	25 317	90 324	838 278	2 792 614
PageRank	14 192	151 993	252 744	2 785 387
PageRank[Δ]	34 507	117 377	437 709	1 869 991
RCR	41 452	124 063	757 046	2 395 072
PRI	68 866	230 227	898 709	3 123 213

Table 5
Significance matrix for the performance comparisons of the single-score metrics based on the high-impact papers (HI). Each table cell shows the confidence levels to which the ranking produced by the metric of the row is significantly better than the ranking produced by the metric in the column. For each table cell the values in the left and right columns correspond to the ACM and MAG databases.

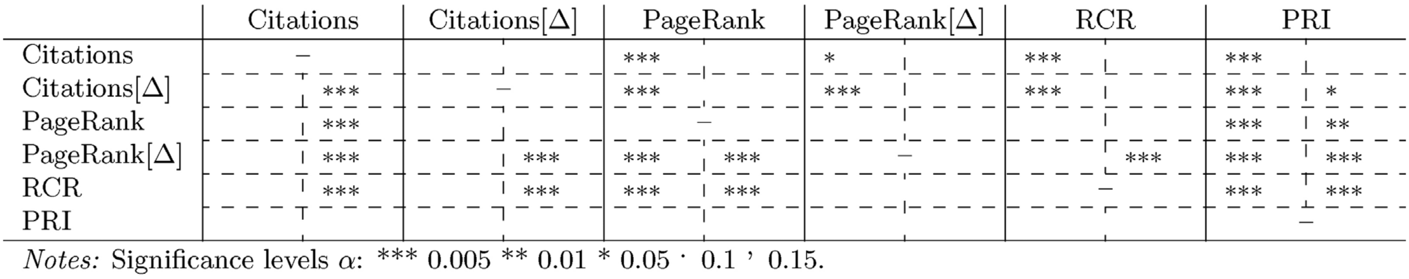


	Citations		Citations[Δ]		PageRank		PageRank[Δ]		RCR		PRI	
Citations	–		***	***			***		***	***	***	***
Citations[Δ]			–				·		***		***	
PageRank		***	*	***	–		***	***	***	***	***	***
PageRank[Δ]		*		***			–			***	***	***
RCR									–		***	*
PRI											–	

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05 · 0.1 ’ 0.15.

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05 · 0.1 ’0.15.

Table 6
Significance matrix for the performance comparisons of the different metrics based on papers that won a best paper award (BPA). Each table cell shows the confidence levels to which the ranking produced by the metric of the row is significantly better than the ranking produced by the metric in the column. For each table cell the values in the left and right columns correspond to the ACM and MAG databases.



	Citations		Citations[Δ]		PageRank		PageRank[Δ]		RCR		PRI	
Citations	–				***		*		***		***	
Citations[Δ]		***	–		***		***		***		***	*
PageRank		***			–						***	**
PageRank[Δ]		***		***	***	***	–			***	***	***
RCR		***		***	***	***			–		***	***
PRI											–	

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05 · 0.1 ’ 0.15.

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05, · 0.1 ’0.15.

approach is the least biased in terms of both time and fields. Relative citation counts and the Abramo method perform similarly well and both show relatively little field bias. After time rescaling ($\Delta = 1000$) citation counts and PageRank their relative biases remain the same, where citation counts are less time biased and PageRank is less field biased. In terms of their field+time bias values, PageRank is slightly less biased than citation counts.

4.2. Performances of single-score metrics

We use the average rank (AR) of the test papers as the evaluation measure to compute performance scores for the metrics that produce a single score per paper. Table 4 lists the metrics' AR values based on the HI and BPA papers for both the ACM and MAG databases. Tables 5 and 6 show the corresponding significance values based on the HI and BPA papers, respectively. For these tables each cell shows the levels to which the ranking produced by the metric of the row is significantly better than the ranking of the metric in the column. The left and right values correspond to the ACM and MAG databases. For these tables and similar tables in the remainder of this paper, we only show significance values in the case where $\alpha < 0.15$.

For example, the last row ‘PRI’ in Table 5 does not contain any significance values which means that the PRI metric does not produce significantly better results compared to any other metric. However, the cell of the first row ‘Citations’ and the fourth column ‘PageRank[Δ]’ contains one value (***) on the left (ACM). This indicates that the difference in the average rank produced by standard citation counts and time-rescaled PageRank is significantly different with 99.5% confidence for the HI data set on the ACM database. In other words, there is a 0.5% probability that the better average rank value produced by citation counts (9933) compared to time-rescaled PageRank's (34 507) is observed by chance.

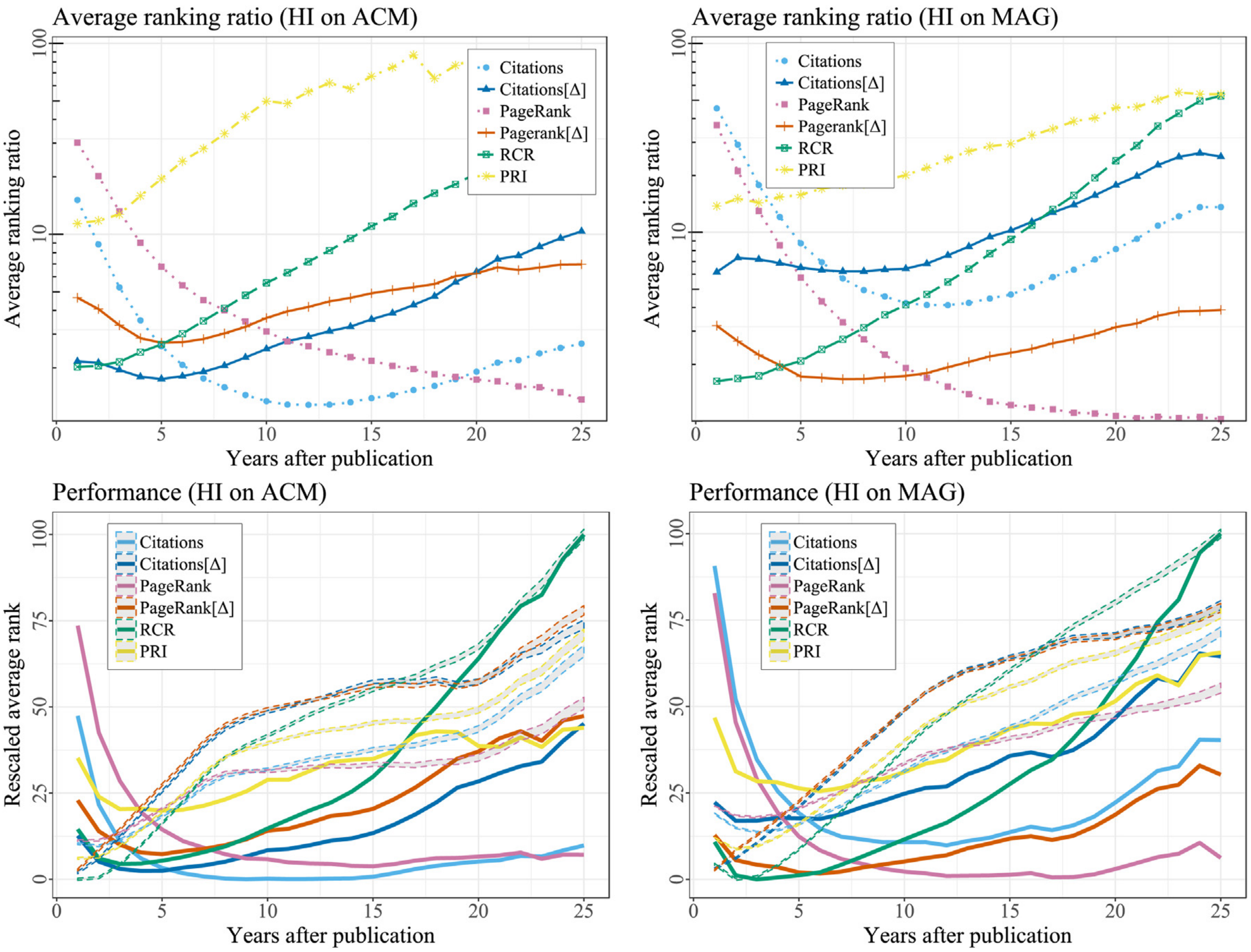


Fig. 4. The average ranking ratio (top) and rescaled average rank (bottom) values for single-score metrics plotted against the time since publication of the HI papers on the ACM (left) and MAG (right) databases. The bottom plots also show the expected average rank values (95% confidence intervals) of the metrics indicating their intrinsic ranking characteristics.

When considering the average ranks of the HI papers on the ACM database, citation counts perform the best (9933) followed by the standard PageRank algorithm (14 192) with no significant difference between their rankings. They perform significantly ($\alpha < 0.005$) better than RCR and PRI. However, their performance reduces significantly ($\alpha < 0.005$) when normalised over time, yet they still perform better than PRI ($\alpha < 0.005$). On the MAG database, standard PageRank performs the best (252 744) followed by its time-rescaled variant (437 709) and standard citation counts (581 642). The PRI metric performs the worst (898 709), however, the difference compared to Citations[Δ] is no longer significant.

In contrast, using the BPA test data set, rescaling citation counts and PageRank over time improves their performances. PageRank improvements are significant ($\alpha < 0.005$) on both databases, while citation counts' improvements are only significant ($\alpha < 0.005$) on the MAG database. For example, PageRank on the MAG database only performs better than RCR when it is time normalised ($\alpha < 0.005$). In general, the PRI metric performs poorly in identifying both the HI and BPA test papers. Apart from the PRI metric, metrics that incorporate time normalisation tend to better identify the BPA papers, while standard citation counts and PageRank tend to rank the HI papers higher.

Since the HI papers are, on average, older than the BPA papers we take the age of the test papers into account in Fig. 4. In other words, we analyse which metrics better identify the HI papers t years after publication. Fig. B.6 shows analogous plots for the BPA test papers. The top two plots in Fig. 4 show the average ranking ratio of the metrics over citation window size t. The time-rescaled citation counts and PageRank identify the HI papers better than their non-normalised counterparts for smaller citation windows. This advantage decreases steadily on both databases with larger citation windows. For citation counts, the cross-over point occurs after 7 years on both databases, while for PageRank it occurs after 9 years (ACM) and 11 years (MAG). This is expected since both citation counts and PageRank are biased towards older papers. Time-rescaling gives both metrics a relative advantage in identifying HI papers shortly after publication. Comparing citation counts to PageRank directly, we find that on the MAG database PageRank always performs better, while on the ACM database PageRank only surpasses citation counts after 20 years. The plot for the MAG database also shows that the RCR metric identifies the test papers the best for the first three years after publication, after which time-rescaled PageRank performs better. The PRI metric performs the worst.

Table 7
The performance (MAR_w) of the metrics that produce a score per category for each paper.

Metric	ACM		MAG	
	MAR_w (HI)	MAR_w (BPA)	MAR_w (HI)	MAR_w (BPA)
Rel. Citations	156	1204	7412	26 800
Abramo	156	1181	7416	26 811
Percentile	184	1302	7271	27 412
Citations[F,Δ]	269	1391	8823	27 172
PageRank[F,Δ]	414	1696	5379	18 153

It is also important to consider the metrics' intrinsic ranking characteristics when evaluated with specific test data. For instance, time-normalised metrics naturally rank younger papers higher than metrics that do not adjust for time. We therefore analyse whether a metric's ranking performance is due to its actual ranking performance or because it naturally ranks the test papers higher due to their age distributions. To simulate a metric's intrinsic ranking characteristics, we use 10 000 samples of papers where each sample comprises randomly selected papers that are distributed according to the publication year distribution of the HI papers. The average ranks produced by the metrics based on these random samples may be interpreted as their expected ranking performance. By comparing the metrics' expected performance trends to their actual performance trends based on the HI papers, we may differentiate for which citation window sizes t metrics perform above their expectations.

In the bottom plots of Fig. 4, we use the HI papers' actual average ranks (AR) for each metric (solid curves) with increasing citation windows since their original publication date. The shaded areas between two dashed curves indicate the expected average ranks (with 95% confidence intervals) of the corresponding metrics, based on the samples of randomly selected papers. We consider the actual average ranks of all metrics and rescale their values to fall between 0 and 100 as follows

$$\text{AR}_{rescaled} = 100 \cdot \frac{\text{AR} - \text{AR}_{\text{min}}}{\text{AR}_{\text{max}} - \text{AR}_{\text{min}}} \tag{6}$$

where AR is the average rank value for a specific metric and citation windows size, while AR_{min} and AR_{max} are the minimum and maximum average rank values for all considered metrics and citation window sizes. We rescale the expected average rank values of all metrics separately. This rescaling is required so that the actual and expected average ranks fit on a common vertical axis. Therefore, the actual ranking trends of the metrics may be compared directly to each other. However, the actual and expected ranking trends are not directly comparable.

We are interested in the relative trends of metrics compared to their expected ranking trends. For example, in both plots the expected trends of time-rescaled citation counts and PageRank are nearly identical with steadily increasing average ranks for larger citation windows. This is expected since both metrics time-rescale paper scores by small reference sets of 1000 similarly aged papers. In contrast, when considering the actual average ranks of the HI papers, the trends between these two metrics are quite different. On the ACM database, Citations[Δ] performs better than PageRank[Δ] for any citation window size. On the MAG database, PageRank[Δ] performs better than Citations[Δ] for any citation window size.

Another example is the RCR metric. The RCR metric is expected to intrinsically perform the best for the first 5 (ACM) and 7 (MAG) years (shaded curves). However, on the ACM database, Citations[Δ] performs similarly well for the first 2 years after which it perform better than RCR. On the MAG database time-rescaled PageRank performs better for citation windows of 6 years and larger. In other words, when considering the expected ranking trends of RCR and PageRank[Δ], then RCR is expected to perform better for $t < 16$ after which PageRank[Δ] is expected to perform better naturally. However, the actual ranking performance cross-over point already occurs at $t = 6$. Therefore, PageRank[Δ] performs better (above expectation) than RCR for citation window sizes of $6 < t < 16$. Following the same argument, we may conclude that PageRank[Δ] on the ACM database performs above expectation compared to the RCR metric for citation window sizes of $9 < t < 16$.

Lastly, we want to highlight the difference between standard citation counts and PageRank. On the ACM database, the expected average rank values for citation counts and PageRank are very similar for the first 10 years, after which PageRank is expected to perform better. However, the actual performance of citation counts in identifying the HI papers is comparatively better for up to 22 years. This indicates that citation counts perform above expectation for $t < 22$ years compared to PageRank. On the MAG database, standard PageRank is expected to perform better than citation counts after 12 years, however, PageRank always outperforms citation counts. These examples show that the HI papers comprise some latent characteristics that are better identified by some metrics compared to others. Furthermore, it should be noted that the metrics' intrinsic ranking characteristics are fairly similar on both databases. However, the actual ranking performances of the metrics differ between the databases. This indicates that the actual ranking performances of the metrics, based on the HI papers, depend on the citation networks of the two databases.

4.3. Performances of metrics with overlapping fields

Table 7 shows the performance of the metrics that normalise over fields and time. Since each paper receives a different score for each field it belongs to, we use the weighted mean average rank (MAR_w) of the test papers as the evaluation measure. In Appendix A we give the definition of MAR_w and discuss the reasons for choosing it as the evaluation measure.

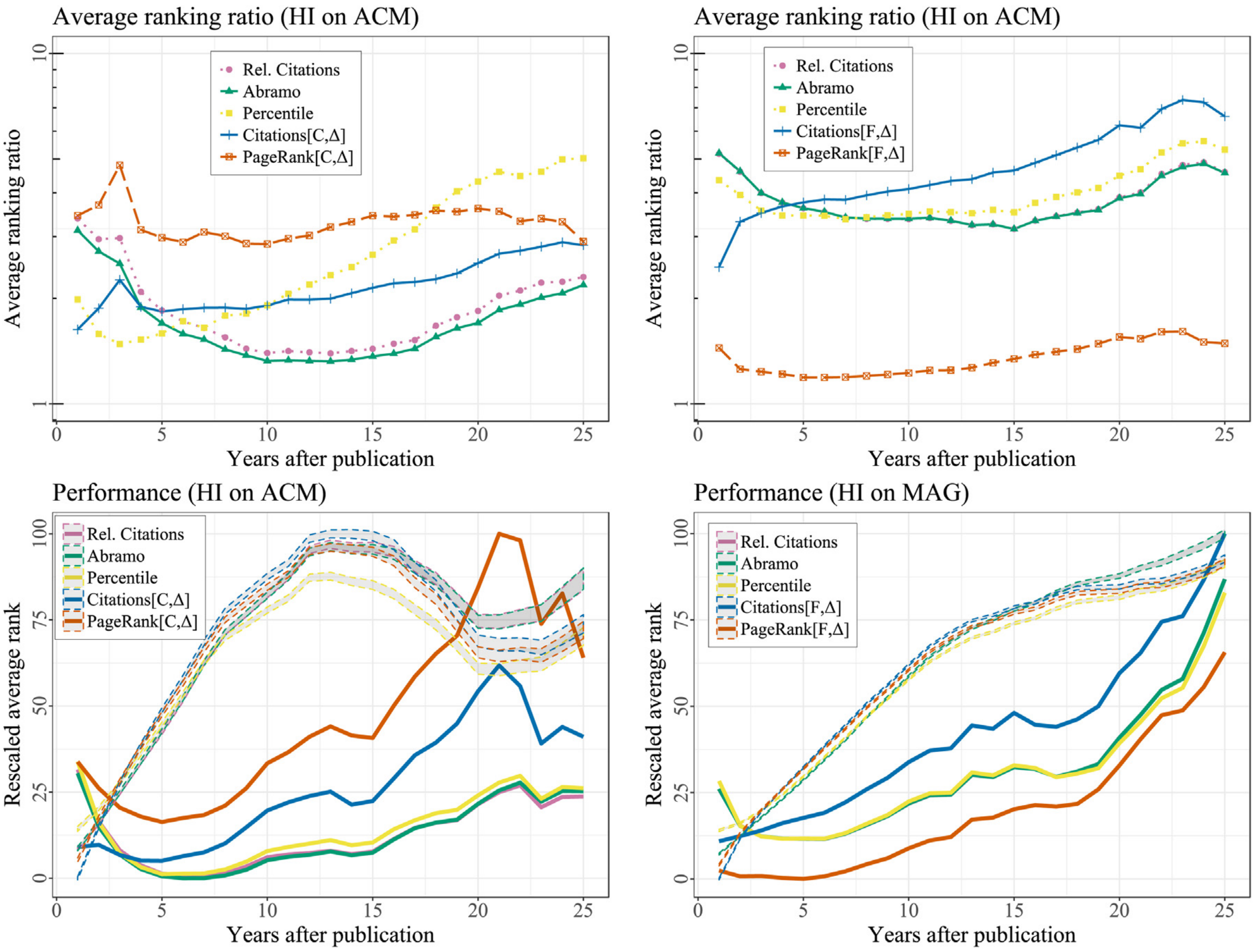


Fig. 5. The average ranking ratio (top) and rescaled average rank values (bottom) for multiple-score metrics plotted against the time since publication of the HI papers on the ACM (left) and MAG (right) databases. The bottom plots also show the expected average ranks values (95% confidence intervals) of the metrics describing their intrinsic ranking characteristics.

On the ACM database, PageRank[F,Δ] performs the worst for both the HI and BPA papers. Compared to PageRank, all other metrics produce significantly ($\alpha < 0.005$) better MAR$_w$ scores. Relative citation and the Abramo method also produce significantly ($\alpha < 0.005$) better rankings than Citations[F,Δ]. Lastly, the percentile approach also improves on Citations[F,Δ] significantly ($\alpha < 0.005$) but only for the HI papers. However, no significant differences are observed between relative citations, the Abramo method, and the percentile approach.

On the MAG database, PageRank[F,Δ] performs the best on both test data sets and achieves significantly ($\alpha < 0.005$) better rankings compared to the other metrics. Using the HI papers, Citations[F,Δ] performs the worst ($\alpha < 0.005$). Again, we find no significant differences between relative citations, the Abramo method, and the percentile approach. Furthermore, using the BPA papers, we find no significant differences between these metrics and Citations[F,Δ].

The top two plots in Fig. 5 show the average ranking ratio of the metrics based on the HI papers with varying citation window sizes. The left and right plots are the results based on the ACM and MAG databases. On the ACM database, PageRank[F,Δ] never performs better in identifying the HI papers compared to Citations[F,Δ]. Citations[F,Δ] ranks the test papers the highest one year after publication but is surpassed by the percentile approach (which normalises over calendar years) up to 5 years after publication. After 5 years, the Abramo method and relative citation counts rank the HI papers the best. The percentile approach steadily performs worse the more years elapse compared to the other metrics. On the MAG database, PageRank[F,Δ] performs the best for any citation window size.

When considering the average rank values of the metrics in the bottom plots in Fig. 5, we find that relative citations, the Abramo method, and the percentile approach perform equally well for any citation window size on both databases. The expected average rank values for these three metrics are also nearly identical for the first 10 years, after which the percentile approach is expected to naturally rank papers slightly better. Time- and field-rescaled PageRank is expected to perform slightly better than Citations[F,Δ] after approximately 5 years. However, independent of citation window size, Citations[F,Δ] performs better on the ACM database, while PageRank[F,Δ] performs better on the MAG database.

4.4. A closer look at PageRank

In this section we briefly describe the performance and bias results of PageRank when it is used with varying parameters. In this section α refers to PageRank's damping factor and not to the significance level. The damping factor has a large influence on the ranking outcome of PageRank because it defines the average path length of the random walkers. It is especially important for paper citation networks due to their time-directed structure. For larger damping values, higher average scores are assigned to older papers (Dunaiski, 2014, p. 97). The influence of publication ages decreases for smaller damping values and when α tends towards 0 all paper scores are roughly the same when no personalisation is used (Chen et al., 2007). Therefore, we varied the damping factor α from 0.05 to 0.95 with 0.05 increments to observe the differences in PageRank's ranking bias and performance.

As expected, we found that PageRank's time bias increases monotonically when α increases. On the ACM database the bias (d_M) ranges from 110.36 ($\alpha = 0.05$) to 195.76 ($\alpha = 0.95$) with a coefficient of variation (CV) of 18.49. Similarly for the MAG database the time bias ranges from 330.62 to 567.55 with a CV of 17.27. When PageRank is time-rescaled, the time bias remains relatively stable with a slight decrease for higher α values. On the ACM (MAG) database the time bias ranges from 12.18 (29.18) to 10.25 (22.62) and a CV of 8.62 (8.21). We also found that the damping factor has very little effect on the field bias of PageRank. On the ACM (MAG) database the CV for the field bias is 3.91 (1.97).

In terms of performance, we found that the optimal α value for PageRank is test data dependent. Standard PageRank performs the best with $\alpha = 0.3$ (HI) and $\alpha = 0.05$ (BPA) on both databases. In general, standard PageRank performs slightly worse with larger α values. When PageRank is time-rescaled, it performs worse with larger α values for the HI papers, while it performs better for the BPA papers. For the HI papers the optimal α values are 0.35 (ACM) and 0.05 (MAG), while for the BPA papers they are 0.85 (ACM) and 0.95 (MAG).

We also considered two additional personalisation approaches to the standard uniform placement of the random walkers. First, we used papers' citation counts as personalisation to see whether forcing the random walkers towards highly cited papers increases PageRank's performance (PageRank(Citations)). Second, we analysed whether using the impact factors of the journals at which papers are published influences the bias and performance of PageRank (PageRank(IF)).

We found that PageRank(IF) reduces time bias but increases field bias compared to standard PageRank. On the ACM (MAG) database time bias reduces by 26.02% (14.72%) while its field bias increases by 36.30% (173.17%). This is expected since the journal impact factor itself is field biased (Althouse, West, Bergstrom, & Bergstrom, 2009). Furthermore, when PageRank is time-rescaled then personalising PageRank with the journal impact factors also increases its time bias by 96.54% and 30.19% on the ACM and MAG database. In general, standard PageRank has the least field bias, while PageRank(Citations) has the least time bias.

In terms of performance, all PageRank variants follow the same trend. The time-rescaled versions only perform better for smaller citation windows. For the MAG database, the cross-over points are between 9 and 11 years for both the high-impact and high-quality papers. On the ACM database, the cross over points are between 8 and 12 years for the high-impact papers, and between 7 and 9 for the high-quality papers. However, we find that PageRank(Citations) performs the best on the ACM database, while standard PageRank performs the best on the MAG database. When the PageRank variants are normalised over time and fields, we observe the same performance results. Again, PageRank(Citations) performs the best on the ACM database, while standard PageRank performs the best on the MAG database for all considered citation window sizes.

5. Discussion

Using a publication database comprising physics papers, Mariani et al. (2016) found that PageRank is more time biased than citation counts. They also showed that after time normalisation, time-rescaled PageRank remains more biased than time-rescaled citation counts. We confirmed these results on both the ACM and MAG databases. We also showed that using $\Delta = 1000$ to rescale citation counts reduces time bias more effectively than normalising over calendar years. In terms of field bias we found that citation counts are always more biased than PageRank. This confirms the results by Vaccario et al. (2017) who used a similar subset of the MAG database. They also showed that PageRank's field bias increases slightly when it is time-rescaled. We obtained similar results on the MAG database. However, we also found that on the ACM database this is not true and that PageRank's concept bias decreases substantially.

Mariani et al. (2016) used a database of 449 935 papers published in journals of the American Physical Society to evaluate citation counts and PageRank on ranking performance. They used test data comprising 87 milestone papers selected for their long-lived contributions to physics, either by announcing significant discoveries, or by initiating new research areas.[1] They found that time-rescaled PageRank identifies these milestone papers the best. Our results showed that, when considering the high-impact papers, time rescaling reduces the overall performance of citation counts and PageRank. This is expected since HI papers are older on average and have had more time to accrue many citations. We also showed that this disadvantage disappears if smaller citation windows are considered and that time-normalised metrics better identify the high-impact papers for the first 7 to 10 years after publication (Fig. 4). Mariani et al. (2016) found that these cross-over points occur after 9 and 18 years for citation counts and PageRank, respectively. They also found that time-rescaled PageRank performs

[1] https://journals.aps.org/prl/50years/milestones.

better than time-rescaled citation counts for any citation window size. Using the MAG database we found the same results. However, on the ACM database, we found that only after 20 years does PageRank[Δ] surpass Citations[Δ] in identifying the HI papers.

When using the BPA test data, we found that time-rescaling improves the performance of both citation counts and PageRank (Table 4). However, we hypothesise that this result will change if the same experiment is executed in a couple of years time with an up to date publication database. We base this statement on the observation that the performance cross-over points between normalised and standard metrics also exist for the BPA papers (see Fig. B.6 in Appendix B). The cross-over point for time-rescaled citation counts to standard citation counts occurs after 6 and 8 years on the ACM and MAG databases. Similarly, for PageRank it occurs after 8 (ACM) and 10 years (MAG). Therefore, when considering the complete databases, the advantage of the standard metrics through their bias towards older papers has not materialised yet since the BPA papers were published more recently on average compared to the HI papers.

This means that time-rescaling should always be implemented unless the objective is to identify high-impact papers published many years ago in which case standard PageRank should be used. Using a test data set of 207 high-impact computer science papers Dunaiski et al. (2016) showed that, with a citation window of 10 years, citation counts perform better on the ACM database while PageRank performs better on the Microsoft Academic Search database (predecessor of MAG). We found the same results for the ACM and MAG databases. However, we cannot confirm that citation counts always perform better when considering complete databases (Dunaiski, 2014; Dunaiski et al., 2016), since we found that PageRank performs better on the MAG database in general. Lastly, when the objective is to find high-quality papers (BPA) published long ago, we found that standard PageRank (MAG database) and standard citation counts (ACM database) perform the best. However, we assume that since the BPA papers are relatively young (average publication year of 2006), the advantage of PageRank has not materialised yet. The cross-over point between standard citation counts and PageRank on the ACM database occurs after 16 years. We therefore hypothesise that shortly after 2022 PageRank will also perform better than citation counts on the ACM database for this set of papers.

When we consider the metrics that normalise for fields and time, then rescaled citation counts are less biased than year-normalised metrics. However, citation counts perform significantly worse. On both databases, when considering different citation windows, rescaled citation counts only perform better for the first one or two years, after which the year-normalised metrics perform better. When factoring in the intrinsic ranking characteristics of the metrics, the performances remain very similar. On the ACM database, the Abramo method and relative citation counts seem to perform slightly better than the percentile approach after approximately 7 to 10 years. On the MAG database rescaled PageRank performs significantly better.

Instead of using predefined field categories, the RCR metric uses papers' co-cited papers to define their fields over which scores are normalised. We found that RCR has more field bias compared to the metrics that normalise over fields. On the ACM database RCR also has more field bias than PRI, which uses journals and conference proceedings as categories. In terms of time bias, only standard citation counts and PageRank exhibit more bias than RCR. This may be explained by the age normalisation of paper scores in the RCR metric, since it depreciates paper scores with age (Janssens et al., 2017). This also explains why RCR performs poorly for larger citation windows (Fig. 4). Lastly, it should be mentioned that the RCR scores are unstable for papers published less than two to three years ago (Hutchins et al., 2016) and for papers with fewer than five citations. Furthermore, Janssens et al. (2017) argue that these thresholds should be chosen even more conservatively.

6. Future work and threats to validity

We assume that the MAG data contains a representative set of papers and citations. Recent studies have shown that the coverage and quality of MAG is comparable to other cross-disciplinary publication databases (Harzing & Alakangas, 2017; Hug & Brändle, 2017). However, it should be noted that the categorisation scheme of MAG for fine-grained topics is too noisy for citation analysis (Hug et al., 2017). We therefore only used the top-level fields in the MAG database in our experiments. Furthermore, we do not draw conclusions about field specific results and therefore the representativeness of an individual field in the MAG database is not important.

Further research is required to directly compare metrics that produce a single score per paper to metrics that produce multiple scores per paper. The problem is to compare a single rank distribution to a set of rank distributions while controlling for field differences. Instead of using the multiplicative counting approach (Albarrán et al., 2011), one could use the approach described by Smolinsky (2016) for overlapping fields in the cases where it is computationally feasible. Whether the results would change remains to be investigated.

We found that PageRank on the MAG database has substantially less field bias compared to the other metrics but not on the ACM database. We assume that since MAG comprises papers from different disciplines, PageRank's citing-side normalisation reduces its field bias. However, PageRank's field bias should be compared to other citing-side normalisation approaches in future investigations.

7. Conclusion

In this paper, we evaluated various article-level impact metrics according to their fairness to rank papers across fields and time, as well as, their performance in identifying high-impact and high-quality papers. We found that inherently PageRank is less field biased while citation counts are less time biased. This does not change when both metrics are normalised over fields and time. When comparing percentile citation scores to mean-normalised citation scores, we found that the percentile approach is less field and time biased. However, we found no significant performance difference between these two metrics.

We showed that time-normalised metrics are better in identifying the high-impact and high-quality papers early. However, for larger citation windows (usually between 5 to 10 years after publication) the standard metrics outperform the time-normalised variants. We also found that PageRank performs the best in identifying the high-impact and high-quality test papers on the MAG database. On the ACM database, citation counts perform better than PageRank. However, for larger citation windows PageRank surpasses citation counts on the ACM database eventually. We also showed that it is important to not only consider different citation window sizes when evaluating the performance of metrics, but to also take their characteristic ranking trends into account. We highlighted a few observations that would otherwise have not been identified.

We found that PageRank's damping factor has a large influence on its time bias, which is substantially reduced when PageRank is time-rescaled. Furthermore, we found that the damping factor has little influence on PageRank's field bias. We also evaluated different PageRank personalisation strategies and found that personalising PageRank with papers' citation counts improves PageRank's time bias. However, PageRank without personalisation has the least field bias. On the MAG database, PageRank without personalisation performs the best, while on the ACM database PageRank, personalised with papers' citation counts, performs the best. Lastly, personalising PageRank with papers' associated journal impact factors results in poor performance and higher time and field bias.

Author contributions

Marcel Dunaiski: Concieved and designed the analysis; Collected the data; Contributed data or analysis tools; Performedthe analysis; Wrote the paper.

Jaco Geldenhuys: Supervision, reviewing, and proof-reading.

Willem Visser: Supervision and reviewing.

Appendix A. The adapted framework for evaluation measures

As mentioned in Section 3.3, we have two performance evaluation scenarios. First, the impact metrics that produce a single score per paper (i.e., a single rank distribution) have to be compared. Second, the field-normalised metrics have to be compared which produce multiple scores for field-overlapping papers. Therefore, multiple rank distributions (one per field) have to be reduced to a single performance score which can be compared between metrics. In this section we describe the evaluation measures used in this paper and the reasons for choosing them.

The average rank (AR) measure simply uses the average rank of the test data entities as performance value. When multiple rank distributions have to be evaluated to compute a performance scores, we define the mean average rank (MAR) evaluation measure

$$\mathrm{MAR} = \frac{1}{|F|} \cdot \sum_{f \in F} \mathrm{AR}(f) \tag{A.1}$$

where F is the set of fields in the database and AR(f) is the average rank of the test papers that belong to field f. Since the number of test papers are not equally distributed among the fields (see Tables B.14 and B.15), the AR value of each field should be weighted proportionally to the number of test papers it contains. Therefore, a weighted variant of MAR may be defined

$$\mathrm{MAR}_w = \frac{1}{|F|} \cdot \sum_{f \in F} \mathrm{AR}(f) \cdot \frac{n_f}{N} \tag{A.2}$$

where n_f is the number of test papers in field f and N is the number of all test papers in all fields (counted multiple times for papers that overlap fields).

Buckley and Voorhees (2000) proposed a framework to compute the stability and sensitivity of evaluation measures in information retrieval problems. An extension of the methodology (Voorhees & Buckley, 2002) can be used to estimate the minimum performance difference required by an evaluation measure to consider two rankings significantly different (at a chosen significance level α). Dunaiski et al. (2018) adapted this framework for rankings of academic entities where the rank distributions of test data are typically skewed and very sparse. We use this framework[2] to identify which evaluation measure should be used for an experiment and to compute the minimum difference in a performance score required to consider two rankings significantly different. Table A.8 depicts the adjusted framework that allows the evaluation of evaluation measures that are computed over sets of queries (multiple rank distributions per metric). In this adapted framework a test collection comprises of C categories (in our case ACM concepts or MAG fields) and q queries for each category. The query $\mathcal{Q}_{i,c}$ comprises the ranks $R(\cdot\,;c)$ of a subset of test papers that belong to category c for query i. In other words, the test paper entities are

[2] The source code is available at: https://github.com/marceldunaiski/RankingEvaluation.

Table A.8
A depiction of a test collection comprising C categories with q different query expressions (i) resulting in unique queries $\mathcal{Q}_{i,c}$ for each category c. A query set $\mathcal{S}_i$ consists of the queries $\mathbf{Q}_{i,c}$, i.e., one query per category.

Category 1			Category 2			···	Category C			Query set
$\mathcal{Q}_{1,1}$	=	$R(\mathrm{P}_1\,;1), R(\mathrm{P}_2\,;1), \ldots$	$\mathcal{Q}_{1,2}$	=	$R(\mathrm{P}_1\,;2), R(\mathrm{P}_2\,;2), \ldots$	···	$\mathcal{Q}_{1,C}$	=	$R(\mathrm{P}_1\,;C), R(\mathrm{P}_2\,;C), \ldots$	$\mathcal{S}_1$
$\mathcal{Q}_{2,1}$	=	$R(\mathrm{P}_{k+1}\,;1), \ldots$	$\mathcal{Q}_{2,2}$	=	$R(\mathrm{P}_{k+1}\,;2), \ldots$	···	$\mathcal{Q}_{2,C}$	=	$R(\mathrm{P}_{k+1}\,;C), \ldots$	$\mathcal{S}_2$
	⋮			⋮		⋱		⋮		⋮
$\mathcal{Q}_{q,1}$	=	$R(\mathrm{P}_{10\cdot k+1}\,;1), \ldots$	$\mathcal{Q}_{q,2}$	=	$R(\mathrm{P}_{10\cdot k+1}\,;2), \ldots$	···	$\mathcal{Q}_{q,C}$	=	$R(\mathrm{P}_{q\cdot k+1}\,;C), \ldots$	$\mathcal{S}_q$

Table A.9
The achieved significance level (ASL) of four evaluation measures (second column) for different significance levels. The best ASL values for each database and test data set combination are highlighted.

Data	Measure	ASL < 0.005	ASL < 0.01	ASL < 0.05	ASL < 0.1	ASL < 0.15
	AR	**87.07**	**92.53**	**98.53**	**99.87**	**100.00**
HI	AR (fields)	50.53	65.87	86.40	91.33	92.80
(ACM)	MAR	61.47	74.80	90.80	92.93	93.47
	MAR_w	53.07	67.73	85.87	91.20	92.67
	AR	**79.07**	**82.67**	**87.87**	**91.87**	**94.67**
BPA	AR (fields)	58.93	65.47	74.53	77.6	79.33
(ACM)	MAR	67.47	71.07	78.00	84.67	87.73
	MAR_w	58.53	65.87	74.67	77.87	79.2
	AR	69.87	75.07	84.4	88.94	92.93
HI	AR (fields)	78.67	**81.73**	89.73	92.53	92.93
(MAG)	MAR	68.27	76.27	**91.07**	**93.2**	**93.33**
	MAR_w	**79.07**	**81.73**	89.73	92.4	92.93
	AR	69.47	75.33	85.6	88.67	90.13
BPA	AR (fields)	**72.53**	75.2	**87.33**	**91.07**	92.4
(MAG)	MAR	69.87	73.47	82.00	86.53	89.87
	MAR_w	72.00	**76.53**	86.53	90.93	**92.8**

Table A.10
The MAR_w scores of the metrics that produce a single score per paper.

Metric	ACM		MAG	
	MAR_w (HI)	MAR_w (BPA)	MAR_w (HI)	MAR_w (BPA)
Citations	118	1329	6051	32 356
Citations[Δ]	325	1535	8876	27 620
PageRank	182	1964	2774	28 244
PageRank[Δ]	467	1882	5094	19 726
RCR	470	1880	8975	26 364
PRI	786	2964	10 031	31 245

split into q different queries for each category where their per-category ranks are used by the evaluation measures. A query set $\mathcal{S}_i$ may contain a paper multiple times, however, since a paper has a different rank for each category it belongs to, each query contains unique rankings.

The *achieved significance level* (ASL) quantifies the sensitivity of an evaluation measures (Sakai, 2006). It is computed by creating $B = 1000$ bootstrap query sets that are sampled with replacement from the q query sets in Table A.8. For each metric pair X, Y and an evaluation measure M, the rate of the number of times that M finds a significant difference (using the two-tailed t-test) between the scores of X and Y is calculated. The more frequently an M's ASL value falls below α, the more sensitive it is to detect differences between lists of rankings. For example, given an ASL < 0.05 and a rate of 80%, we can be 95% confident that M identifies differences in rankings 80% of the time.

In Section 4.2 we listed the performance results of the single-score metrics using AR as the evaluation measure without considering the test papers' fields. We show here that the performance results of the metrics do not change significantly if we consider the test papers' fields. Table A.9 shows the ASL values at different significance levels for the three evaluation measures with field-dependent queries (AR (fields), MAR, MAR_w) and the AR measure without considering fields separately. On the ACM database, the AR without field distinctions has the highest ASL values for all α for both the HI and BPA test data sets. Of the three field-dependent metrics, MAR performs the best while AR (fields) and MAR_w performs very similar. On the MAG database no clear best measure exists, however, the ASL values for the measures are fairly similar for all significance levels.

This indicates that the average rank without field distinctions is the preferred evaluation measure since it is slightly more sensitive to discriminate between the metrics considered in Section 4.2. However, in Table A.10 we show the metrics' MAR_w scores to analyse how the results change compared to simply using the AR values. Tables A.11 and A.12 report the

Table A.11
Significance matrix for the performance comparisons of the single-score metrics based on the weighted average rank (MAR$_w$) of the HI papers. It reports the significance levels to which the metric of the row performs significantly better than the metric of the column. The left and right values in each table cell correspond to the ACM and MAG databases.

	Citations		Citations[Δ]		PageRank		PageRank[Δ]		RCR		PRI	
Citations	–		*	***			***		***	***	***	***
Citations[Δ]			–								***	*
PageRank		***		***	–		***	***	***	***	***	***
PageRank[Δ]		·		***			–			***	***	***
RCR									–		***	*
PRI											–	

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05 · 0.1 ’0.15.

Table A.12
Significance matrix for the performance comparisons of the single-score metrics based on the weighted average rank (MAR$_w$) of the BPA papers. It reports the significance levels to which the metric of the row performs significantly better than the metric of the column. The left and right values in each table cell correspond to the ACM and MAG databases.

	Citations		Citations[Δ]		PageRank		PageRank[Δ]		RCR		PRI	
Citations	–				***		***		***		***	
Citations[Δ]		***	–		***		*		*		***	***
PageRank		***			–						***	***
PageRank[Δ]		***		***		***	–			***	***	***
RCR		***		,		**			–		***	***
PRI											–	

Notes: Significance levels α: *** 0.005 ** 0.01 * 0.05 · 0.1 ’0.15.

Table A.13
The achieved significance level (ASL) of the three field-dependent evaluation measures (second column) for different significance levels. The best ASL values for each database and test data set combination are highlighted.

Data	Measure	ASL < 0.005	ASL < 0.01	ASL < 0.05	ASL < 0.1	ASL < 0.15
HI (ACM)	AR (fields)	42.80	59.20	85.60	89.00	**90.00**
	MAR	42.60	61.60	85.20	**89.60**	**90.00**
	MAR$_w$	**54.40**	**67.60**	**86.60**	89.20	89.80
BPA (ACM)	AR (fields)	**28.00**	36.40	**76.40**	**94.40**	96.80
	MAR	26.40	36.80	**76.40**	93.20	**97.20**
	MAR$_w$	25.60	**38.20**	75.60	90.40	96.00
HI (MAG)	AR (fields)	**68.00**	77.80	96.20	98.80	**99.60**
	MAR	66.40	**79.00**	**96.60**	**99.20**	**99.60**
	MAR$_w$	56.00	66.40	78.60	80.00	80.00
BPA (MAG)	AR (fields)	**63.60**	**67.60**	69.80	**72.00**	77.00
	MAR	**63.60**	66.80	**70.00**	**72.00**	**77.20**
	MAR$_w$	46.80	48.20	53.00	62.60	68.60

corresponding significance values for the performance differences of the metrics based on the HI and BPA test data sets, respectively.

The performance results (MAR$_w$) of the metrics while taking the fields into consideration are very similar to the results when total ranks are used for evaluation (Table 4). A few differences exist. On the ACM database and using the BPA papers, time-rescaled citation counts performs worse than standard citation counts when evaluated with MAR$_w$. However, the performance differences are not significant based on either AR (Table 6) or MAR$_w$ (Table A.12). The same holds true for all other occurrences where the performance orders switch around: PageRank[Δ] and RCR (BPA on ACM); Citations[Δ] and RCR (HI on MAG); and Citations[Δ] and PageRank (BPA on MAG). Apart from these differences, the level of a few significance values change. This shows that the results remain relatively consistent, irrespective of whether total ranks or per-field ranks are used for evaluation.

Table A.13 shows the ASL values at different significance levels for the evaluation measures that use multiple queries per metric. In Section 4.3 we used the weighted mean average rank (MAR$_w$) as evaluation measure for the metrics that assign multiple scores per paper. We chose MAR$_w$ since it makes more sense formally that fields with more test data entities should count more. Furthermore, the different evaluation measures in Table A.13 have relatively similar ASL values.

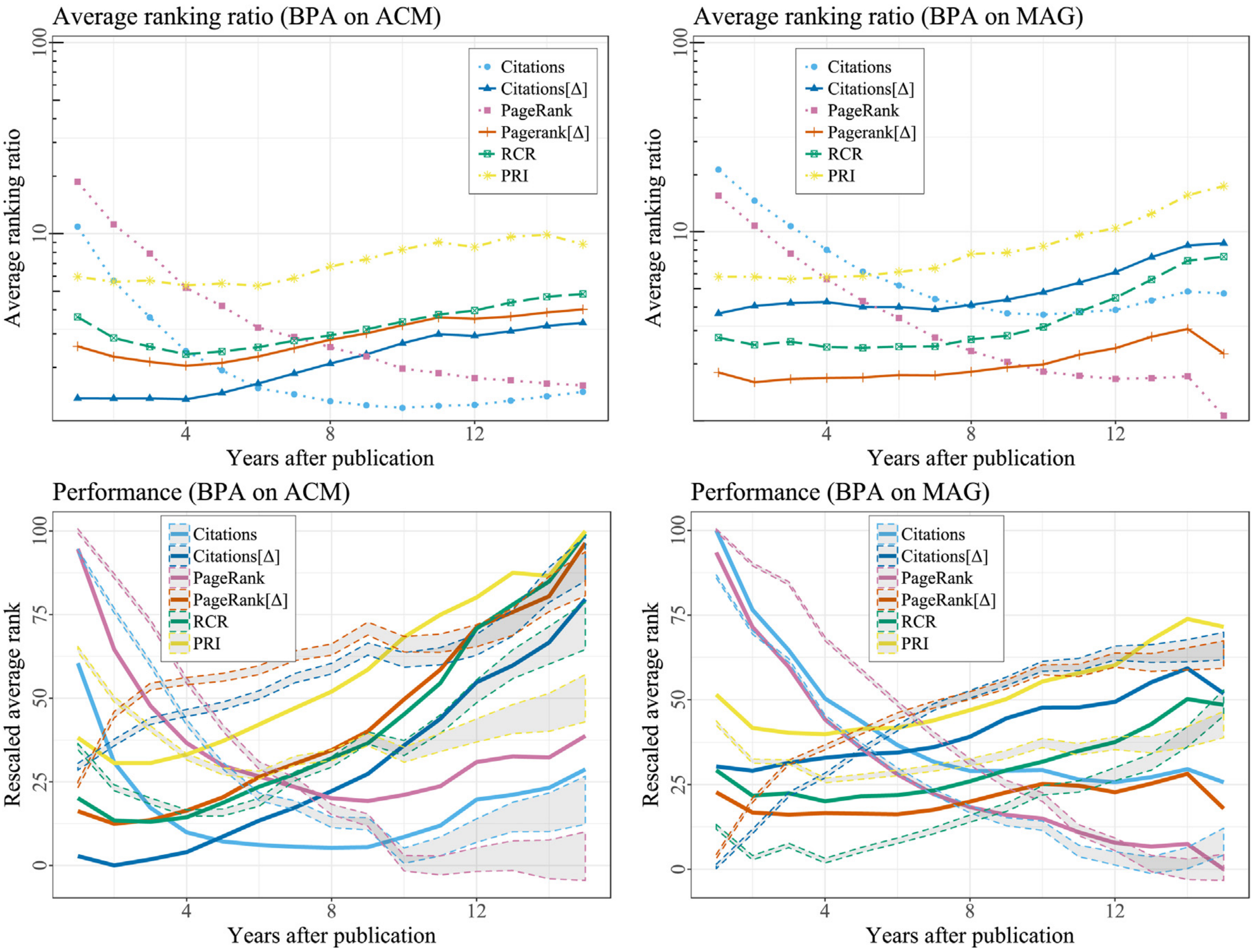


Fig. B.6. The average ranking ratio (top) and rescaled average rank (bottom) values for single-score metrics plotted against the time since publication of the BPA papers on the ACM (left) and MAG (right) databases. The bottom plots also show the expected average rank values (95% confidence intervals) of the metrics indicating their intrinsic ranking characteristics.

Appendix B. Supplementary information

Compared to Mariani et al. (2016, Equation F1) we only use time intervals of a whole calendar year which simplifies the formula for the average ranking ratio. Let $t^{(c)}$ denote the calendar year for which all papers published until $t^{(c)}$ are ranked by the considered ranking metrics. The average ranking ratio $\bar{r}(m, t)$ for metric m and citation window size t is defined as follows:

$$\bar{r}(m,t) = \frac{1}{M(t)} \sum_{t^{(c)}} \sum_{i \in M} \delta\left(t^{(c)} - t_i, t\right) \times \frac{r_i(m, t^{(c)})}{\min_{m'}\{r_i(m', t^{(c)})\}} \tag{B.1}$$

where t_i is the publication year of test paper i, $r_i(m, t^{(c)})$ is the rank of i at year $t^{(c)}$ according to metric m, $M(t)$ is the number of test papers that are at least t years old, and $\delta(x, y)$ denotes the Kronecker delta function of x and y which evaluates to 1 if x is equal to y and 0 otherwise.

The top two plots in Fig. B.6 show the average ranking ratio of the single-score metrics evaluated in Section 4.2 based on the BPA test data sets. Analogous to Fig. 4 we vary the citation window size t and include the expected ranking trends in bottom two plots. We are interested in whether the BPA papers exhibit some latent characteristic that distinguishes them the average paper. Therefore, the metrics' expected ranking trends should be compared to their actual ranking trends.

In other words, when comparing two metrics and their expected rankings trends are identical to their actual ranking trends based on the BPA papers, then we may conclude that either metric is unable to discriminate between the BPA papers and sets of randomly selected papers. For example, when comparing the expected ranking trends of standard citation counts and time-rescaled citation counts, the cross-over points occur at $t = 4$ (ACM) and $t = 5$ (MAG). The cross-over points of the actual ranking trends occur at $t = 5$ (ACM) and $t = 6$ (MAG). The cross-over points of the expected ranking trends are only shifted forward by one year compared the actual ranking trends. This indicates that relative to each other, these two metrics identify the latent characteristics of the BPA papers poorly. However, some measures perform different to their expectations. For example, when comparing citation counts to PageRank on the ACM database, we observe that the expected cross-over

Table B.14
The top-level concepts of the ACM database with their number of papers and average citation counts. Papers are counted multiple times since a paper is assigned to one or more top-level concepts based on the concepts that are most frequently assigned to it. Columns 'HI papers' and 'BPA papers' show the number of test papers per concept. The last row reports the total number of unique papers.

Concept	Papers	Average citations	HI papers	BPA papers
Applied computing	98 794	2.85	7	5
Computer systems organization	31 583	6.30	17	7
Computing methodologies	156 981	6.56	30	44
General and reference	33 237	2.42	12	11
Hardware	50 400	4.35	14	3
Human-centered computing	43 369	6.30	3	67
Information systems	90 588	7.30	63	87
Mathematics of computing	127 090	4.16	24	21
Networks	46 344	5.98	33	21
Security and privacy	14 230	7.05	5	5
Social and professional topics	54 319	3.36	10	16
Software and its engineering	92 507	6.99	133	130
Theory of computation	84 281	7.35	48	51
None	1 033 885	2.45	85	136
Total papers	1 957 608	–	484	604
Unique papers	1 737 687	–	401	516

Table B.15
The top-level fields of the MAG database and their number of papers with average citation counts. Papers are counted multiple times since paper may be field-overlapping. Columns 'HI papers' and 'BPA papers' show the number of test papers per field. The last row reports the total number of unique papers.

Field	Papers	Average citations	HI papers	BPA papers
Art	137 526	10.54	14	25
Biology	4 935 116	19.05	17	30
Business	423 171	10.90	21	29
Chemistry	4 975 005	14.98	8	9
Computer Science	3 002 769	13.56	337	438
Economics	1 636 621	12.27	78	128
Engineering	2 122 686	11.54	99	167
Environmental science	240 439	26.42	0	3
Geography	220 697	14.31	2	8
Geology	1 448 707	15.90	5	9
History	277 970	12.65	4	16
Materials Science	1 563 279	13.35	2	11
Mathematics	3 533 736	12.20	201	242
Medicine	4 118 277	16.87	10	34
Philosophy	565 568	11.48	41	72
Physics	5 172 391	12.08	53	79
Political Science	80 126	7.29	2	3
Psychology	2 323 402	16.76	25	102
Sociology	1 303 427	12.06	55	122
Total papers	38 080 913	–	974	1 527
Unique papers	13 829 901	–	354	505

point occurs at $t = 10$. In contrast, the actual ranking trends have no cross-over point and citation counts identify the BPA papers better than PageRank for all considered citation windows. From this observation we may conclude that some latent characteristic of the BPA papers exists which is better identified by citation counts than PageRank on the ACM database.

References

ACM Inc. (2014). *ACM Digital Library*.. . [Online]; Accessed 12.09.18. http://dl.acm.org/

ACM Inc. (2017). *The 2012 ACM Computing Classification System*.. . [Online]; Accessed 12.09.18. http://www.acm.org/publications/class-2012-intro

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom – A test of Zitt's hypothesis. *Scientometrics*, *75*(1), 81–95.

Albarrán, P., Crespo, J. A., Ortu no, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385–397.

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2009). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, *60*(1), 27–34.

Bornmann, L. (2017). Measuring impact in research evaluations: A thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, *73*(5), 775–787.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, *7*(1), 158–165.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418.

Bornmann, L., & Marx, W. (2018). Critical rationalism and the search for standard (field-normalized) indicators in bibliometrics. *Journal of Informetrics*, *12*(3), 598–604.

Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, *66*(11), 2215–2222.

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).

Buckley, C., & Voorhees, E. M. (2000). Evaluating evaluation measure stability. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'00* (pp. 33–40).

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

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, *5*(1), 101–113.

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

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). How to evaluate rankings of academic entities using test data. *Journal of Informetrics*, *12*(3), 631–655.

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).

Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, *10*(2), 392–407.

Giuffrida, C., Abramo, G., & D'Angelo, C. A. (2018). *Do all citations value the same? Valuing citations by the value of the citing items.* , https://arxiv.org/abs/1809.06088arXiv:1809.06088.

Gringorten, I. I. (1963). A plotting rule for extreme probability paper. *Journal of Geophysical Research*, *68*(3), 813–814.

Harzing, A.-W., & Alakangas, S. (2017). Microsoft academic: Is the phoenix getting wings? *Scientometrics*, *110*(1), 371–383.

Herranz, N., & Ruiz-Castillo, J. (2012). Sub-field normalization in the multiplicative case: High- and low-impact citation indicators. *Research Evaluation*, *21*(2), 113–125.

Hug, S. E., & Brändle, M. P. (2017). The coverage of microsoft academic: Analyzing the publication output of a university. *Scientometrics*, *113*(3), 1551–1571.

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with microsoft academic. *Scientometrics*, *111*(1), 371–378.

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016). Relative citation ratio (RCR): A new metric that uses citation rates to measure influence at the article level. *PLoS Biology*, *14*(9), e1002541.

Ioannidis, J. P. A., Boyack, K., & Wouters, P. F. (2016). Citation metrics: A primer on how (not) to normalize. *PLoS Biology*, *14*(9), e1002542.

Janssens, A., Goodman, M., Powell, K., & Gwinn, M. (2017). A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). *PLoS Biology*, *15*(10), e2002536.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics*, *1*(2), 145–154.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the national institute of science of India*, 49–55.

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*(4), 1207–1223.

Microsoft. (2017). *Microsoft academic graph*. [Online]; Accessed 15.08.17. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001.

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*(4), 734–745.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, *12*(5), 297–312.

Pudovkin, A., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. *Collnet Journal of Scientometrics and Information Management*, *3*(2), 3–10.

Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics*, *6*(1), 121–130.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17268–17272.

Sakai, T. (2006). Evaluating evaluation metrics based on the bootstrap. *Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval*, 525–532.

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

Silva, F. N., Rodrigues, F. A., Oliveira, O. N., da, F., & Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields. *Journal of Informetrics*, *7*(2), 469–477.

Sirtes, D. (2012). Finding the easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*(3), 448–450.

Smolinsky, L. (2016). Expected number of citations and the crown indicator. *Journal of Informetrics*, *10*(1), 43–47.

Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics*, *11*(3), 766–782.

van Leeuwen, T. N., & Calero Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation*, *21*(1), 61–70.

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR'02* (pp. 316–323).

Waltman, L. (2015). *NIH's new citation metric: A step forward in quantifying scientific impact?* [Online]; Accessed 10.09.18. https://www.cwts.nl/blog?article=n-q2u294

Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: An empirical analysis. *Scientometrics*, *87*(3), 467–481.

Waltman, L., Yan, E., & van Eck, N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics*, *89*(1), 301.

Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*(1), 72–77.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401.

Journal of Informetrics 13 (2019) 299–313

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Globalised vs averaged: Bias and ranking performance on the author level

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[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 12 November 2018
Received in revised form 12 January 2019
Accepted 12 January 2019

Keywords:
Author impact indicators
Globalised vs. averaged
Field bias
Ranking evaluation
Test data

ABSTRACT

We analyse the difference between the averaged (average of ratios) and globalised (ratio of averages) author-level aggregation approaches based on various paper-level metrics. We evaluate the aggregation variants in terms of (1) their field bias on the author-level and (2) their ranking performance based on test data that comprises researchers that have received fellowship status or won prestigious awards for their long-lasting and high-impact research contributions to their fields. We consider various direct and indirect paper-level metrics with different normalisation approaches (mean-based, percentile-based, co-citation-based) and focus on the bias and performance differences between the two aggregation variants of each metric. We execute all experiments on two publication databases which use different field categorisation schemes. The first uses author-chosen concept categories and covers the computer science literature. The second covers all disciplines and categorises papers by keywords based on their contents. In terms of bias, we find relatively little difference between the averaged and globalised variants. For mean-normalised citation counts we find no significant difference between the two approaches. However, the percentile-based metric shows less bias with the globalised approach, except for citation windows smaller than four years. On the multi-disciplinary database, PageRank has the overall least bias but shows no significant difference between the two aggregation variants. The averaged variants of most metrics have less bias for small citation windows. For larger citation windows the differences are smaller and are mostly insignificant.

In terms of ranking the well-established researchers who have received accolades for their high-impact contributions, we find that the globalised variant of the percentile-based metric performs better. Again we find no significant differences between the globalised and averaged variants based on citation counts and PageRank scores.

1. Introduction

Citation metrics constitute a key tool in scientometrics and play an increasingly important role in the evaluation of researchers (Bornmann, 2017). To enable fair evaluations, it is a de facto requirement that metrics are field and time normalised (Waltman, 2016). On the paper level, a paper's actual score is usually compared to the expected score computed

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

https://doi.org/10.1016/j.joi.2019.01.006
1751-1577/© 2019 Elsevier Ltd. All rights reserved.

from a reference set comprising papers from the same field and published in the same year. Normalised paper impact scores may then be aggregated to define author-level impact metrics. Two aggregation approaches exist that use the actual and expected scores of papers. The first computes the average of each paper's ratio of actual and expected scores, which is often referred to as the 'average of ratios' or *averaged* approach (Waltman, 2016). The second divides the sum of an author's actual paper scores by the sum of the corresponding expected paper scores. The latter is also referred to as the 'ratio of averages' or *globalised* approach (Egghe & Rousseau, 1996). Opinions differ as to which approach is better suited or more appropriate for the evaluation of academic entities (Egghe & Rousseau, 1996; Lundberg, 2007; Moed, 2010; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a). In this paper we take a quantitative look at the differences between these two aggregation approaches.

Since no gold standard for normalised metrics exists (Bornmann & Marx, 2018), we use a number of different paper-level impact metrics that use different normalisation strategies to overcome the bias introduced through varying citation potentials between research fields and time. For instance, we use mean-normalised citation scores where a paper's citation count is compared to the mean (expected) citation count of papers published in its field and in the same year (Lundberg, 2007; Radicchi, Fortunato, & Castellano, 2008). We also use a percentile metric where a paper is rated in terms of its percentile in the score distribution of papers in the same field and with the same publication year (Bornmann, Leydesdorff, & Mutz, 2013; Leydesdorff, Bornmann, Mutz, & Opthof, 2011). We also look at indirect metrics (Giuffrida, Abramo, & D'Angelo, 2018; Pinski & Narin, 1976) and a metric where a paper's co-cited papers (papers that are cited together with it) are used as the reference set to compute expected paper scores (Hutchins, Yuan, Anderson, & Santangelo, 2016).

An important task in scientometrics is to validate that metrics fulfil their intended purpose. According to Bornmann and Marx (2018), situations should be created or found in empirical research in which a metric can fail to achieve its purpose. A metric should only be regarded as provisionally valid if these situations could not be found or realised. For example, metrics that are intended to rate the quality of papers should be assessed by correlating them with peer assessments (Bornmann & Marx, 2015). This also applies to author-level evaluations. However, collecting direct peer-assessed test data is time consuming and expensive. We therefore use a proxy for this assessment which comprises test data based on awards and other recognitions that researchers have received for their outstanding contributions in their fields (Dunaiski, Geldenhuys, & Visser, 2018a; Dunaiski, Visser, & Geldenhuys, 2016; Fiala, Šubelj, Žitnik, & Bajec, 2015; Fiala, 2012; Fiala, Rousselot, & Ježek, 2008; Fiala & Tutoky, 2017; Gao, Wang, Li, Zhang, & Zeng, 2016; Nykl, Campr, & Ježek, 2015; Nykl, Ježek, Fiala, & Dostal, 2014). Specifically, we use selected researchers that have won prizes for their highly influential and long-lasting contributions and researchers that have been awarded the ACM fellowship for similar achievements.

We follow the appeal by Bornmann and Marx (2018) for continued scrutiny of current proposals and analyse the difference between the averaged and globalised variants of various paper-level metrics along two dimensions: (1) their fairness to rank authors across fields, and (2) their performance in ranking the well-established researchers comprising our test data set. We compare the overall bias and performance of the metrics but focus on the differences between the averaged and globalised variants for each paper-level metric.

We conduct all experiments on two publication databases. The first database is the ACM Digital Library (ACM, Inc, 2014) which provides a Computing Classification System (CCS) that consists of a library-like, hierarchical structure of concepts. Authors may assign their papers to one or more concepts in this classification hierarchy. We use the CCS to categorise papers and authors into subfields of the computer science discipline. The second database is the Microsoft Academic Graph (MAG) database (Microsoft, 2017). It is multi-disciplinary and papers are assigned to fields in a hierarchical structure based on keywords extracted from their texts. We use the top-level fields as paper categories which roughly capture the scientific disciplines such as 'Mathematics' or 'Medicine'. Again, we categorise authors into disciplines based on their published work.

With this paper we make the following contributions:

- We analyse the averaged and globalised aggregation approaches on the author-level using two different field classification schemes. The first is a categorisation where the authors chose their papers' categories (ACM database). The second is based on semantic information contained within titles and abstracts (MAG database).
- We consider a range of paper-level metrics and show that for some metrics the choice between using the averaged or the globalised approach is important and impacts the author-level metric's field bias as well as its performance in identifying well-established researchers.
- We analyse the bias and performance of the variants over a range of citation window sizes (1–25 years). We find that the choice between the aggregation variants depends less on citation windows sizes. However, the differences between metrics change substantially with different citation windows.

In this paper we first provide the reader with background information about normalisation factors and focus on the arguments for or against the averaged and globalised approaches (Section 2). In Section 3, we describe the methodology of evaluating the metrics along the bias and performance dimensions. We present the results in Section 4, followed by a discussion of the results in Section 5.

2. Background information

2.1. Paper-level normalisation

One of the key principles of bibliometrics is that entities from different fields should not be compared directly based on total citations counts. This stems from the observation that citation densities (mean citation counts) vary between fields due to their different sizes and publication cultures (Lundberg, 2007; Radicchi et al., 2008). Citation densities may even vary between narrow subfields within the same discipline (van Leeuwen & Calero Medina, 2012). In addition, citation counts of papers with different publication years should also not be used directly for comparisons since older papers have had more time to accrue citations. For example, a paper published in 2000 with 40 citations should not necessarily be considered more impactful than a paper from 2010 with 10 citations, if the average citation count per paper in 2000 was 80 while in 2010 it was 10.

To overcome these biases, mean-based metrics have been suggested that normalise papers' citation counts by the average citation count in a field and year (Radicchi et al., 2008). However, citation distributions of papers are inherently skewed, with many papers that only obtain a few citation or none at all, and long-tailed (only a few papers receive a large number of citations). Therefore, percentile-based metrics have been proposed where the citation score of a paper is rated in terms of its percentile in the citation distribution of the field and year to which it belongs (Bornmann et al., 2013; Leydesdorff et al., 2011).

The difficulty in normalising for fields is how to assign papers to fields. In the past, fields have been categorised on the basis of journals or library categories. The problem is that fields and especially broad disciplines are not isolated. Generally, within-field citations are denser than between-field citations, however, between-field citations are becoming more common nowadays (Silva, Rodrigues, Oliveira, da, & Costa, 2013). For example, it has been shown that by considering only the highest level of aggregation (i.e., disciplines), heterogeneities in the subfields' citation patterns might be disregarded (van Leeuwen & Calero Medina, 2012). This is more problematic for indirect metrics (such as PageRank) compared to citation counts (Waltman, Yan, & van Eck, 2011). Agreement about the optimal classification scheme has yet to be reached (Adams, Gurney, & Jackson, 2008; Colliander & Ahlgren, 2011; Zitt, Ramanana-Rahary, & Bassecoulard, 2005).

Alternatives to defining fields a priori exist. Fields may be defined algorithmically based on the citation network structure (Rosvall & Bergstrom, 2008) or by clustering papers into a hierarchical classification system based on direct citation relations between papers (Ruiz-Castillo & Waltman, 2015; Waltman & van Eck, 2012). Furthermore, the semantic relatedness of papers' contents may be leveraged for weighting the similarity between papers (Colliander, 2015). A paper's co-citation network may also be used to define its field (Hutchins et al., 2016). The rationale is that if papers that are cited together by another paper, they belong to the same topic since they were relevant in producing the new paper. However, this does not hold for all citations. Janssens, Goodman, Powell, and Gwinn (2017) found that most of the co-cited papers, that are in a paper's co-citation network due to only a few co-citations, do not belong to the same topic. Furthermore, since some work is inherently of such a nature that it attracts citations from multiple, generally unrelated fields, such as statistical methods (Silva et al., 2013).

Normalising over ill-defined fields may also lead to undesirable situations. For example, when a paper's field is defined by its co-citations and it receives new citations from a remote field it may indicate an increase in importance. However, as Waltman (2015) points out, if the remote field has a high citation density, normalisation may lead to a decrease rather than an increase in the paper's score with the newly acquired citations. In this paper, we use two different categorisation schemes. On the ACM database, we use the ACM classification system (ACM, Inc, 2017b) which consists of library-like categories that are fine-grained and author-chosen. We also use the MAG field categorisation scheme which is based on the semantic information contained in keywords, titles, and abstracts (Microsoft, 2017).

2.2. Aggregating to the author level

Author impact scores may be computed as an aggregation of associated paper impact scores. Alternatively, author impact metrics may be directly defined for author co-citation graphs (Ding, Yan, Frazho, & Caverlee, 2009; West, Jensen, Dandrea, Gordon, & Bergstrom, 2013) or author collaboration graphs (Fiala et al., 2008; Liu, Bollen, Nelson, & Van de Sompel, 2005). In this paper, we only consider the first type of metrics and focus on the different approaches of aggregating paper scores for authors. Higher level aggregation approaches are similar, however, in this paper we only focus on the author-level aggregation since our test data only comprises author entities. Table 1 shows an illustrative example where five papers P1 through P5 are associated with authors A1 through A5. The third column lists the papers' citation counts. Each paper may belong to one or more fields (column four). In the case of mean-normalised metrics, a paper's normalised score is the ratio of its actual score (Citations) and its expected score (average citation counts of papers from the same field and year). To compute the expected scores of fields, a paper and its score is divided equally between fields if it belongs to more than one field. Furthermore, the expected score of a field-overlapping paper is the harmonic mean of its fields' expected scores (Waltman et al., 2011a).

For example, the expected score of papers that only belong to field X (P1 and P5) is $\frac{12+6/2+2}{1+1/2+1} = 6.80$. Similarly, for papers that exclusively belong to field Y (P2) the expected score is $\frac{1+6/2}{1+1/2} = 2.67$. For paper P4, which belongs both to field X and Y,

Table 1
Example of the calculation of expected and normalised citation scores for a set of papers.

Paper	Authors	Citations	Fields	Expected Score	Normalised Score
P1	A1	12	X	6.80	1.76
P2	A2 and A3	1	Y	2.67	0.38
P3	A4	6	Z	6.00	1.00
P4	A1 and A3	6	X and Y	3.83	1.57
P5	A1 and A5	2	X	6.80	0.29

Table 2
Example of the calculation of the three aggregation approaches for a set of authors.

Authors	Total	Averaged	Globalised
A1	3.63	1.21	1.15
A2	0.38	0.38	0.38
A3	1.94	0.97	1.08
A4	1.00	1.00	1.00
A5	0.29	0.29	0.29

the expected score is the harmonic mean of the expected scores for the fields, i.e., $\frac{2}{1/6.80+1/2.67} = 3.83$. It should be noted that since the sum of all paper scores between fields X and Y may not be equal, P4's score should ideally be attributed to field X and Y proportionally (Smolinsky, 2016). Alternatively, if the score of P4 is evenly attributed to fields X and Y, then the paper has to be attributed to the fields proportionally since the number of papers in fields X and Y may not be equal. However, these two alternative approaches are computationally very demanding because it requires to solve systems of non-linear equations (the size of all field combinations) for each year. We therefore use the simpler approach in which papers and scores are equally divided between overlapping fields.

The simplest approach to aggregate paper scores to an author is to simply compute the sum of all associated normalised paper scores. This yields a size-dependent metric variant (total) where author scores cannot decrease with newly added papers. Alternatively, size-independent variants may be defined by using the papers' actual and expected citation scores. The **globalised** approach uses the ratio of the sum of papers' actual scores and the sum of the papers' expected scores (Egghe & Rousseau, 1996). The **averaged** approach uses the average of the normalised paper scores (i.e., the papers' individual ratios). Table 2 depicts these different approaches. Consider author A1 who has three papers P1, P4 and P5. The total (i.e., size-dependent) approach is the sum of the normalised paper scores, i.e., 1.76 + 1.57 + 0.29 = 3.63. The averaged approach is the average of the each paper's ratio: $\frac{1}{3}(12/6.80 + 6/3.83 + 2/6.80) = 1.21$. Lastly, the globalised approach is the ratio of the sum of the actual scores and the sum of the expected scores, i.e., (12 + 6 +2)/(6.80 + 3.83 + 6.80) = 1.15.

The main difference between these approaches is that with the globalised variant an entity's papers are considered an indiscriminate oeuvre rather than a set of individual papers (Waltman et al., 2011a). With this notion, the citation distribution of the entity's papers is not considered important and only the total citation count received by the oeuvre is considered important. Moed (2010) agues that this notion is judicious, at least for the level of research groups, since they tend to produce a coherent set of papers and citing authors do not distribute citations evenly among the group's oeuvre. Instead, only a small set of 'flag' papers are cited which should be seen as citations to the entire oeuvre. The same argument may be made for the author level, since authors tend to produce work in coherent sets of topics matching their expertise. Furthermore, since authors build on their previous work, citing authors will not always distribute citations over all of the authors' previous papers. The globalised approach simply detaches the citations from the author's papers which actually receive them. Consequently, an author's total number of citations is compared to the expected number of citations of a set of papers with the same size, and the same distribution across subject fields and years. The globalised approach is therefore an author's normalised impact of their published work, while the averaged approach is the authors' average normalised impact per paper.

Vinkler (2012) argues along similar lines and that the averaged approach is not well-suited for the application of scientometric indicators for research evaluation (team or institute level). For example, he points out that when a paper with an expected citation score below 1 receives one or a small number of citations, the averaged approach distorts values significantly. On the author level, this approach therefore may unfairly bias authors with a relatively low number of papers or citations. Lastly, he finds that the citation distribution in different fields, the sizes of the fields, and the number of fields taken into consideration also influences results based on averaged citation scores.

In contrast, Waltman et al. (2011a) argue for the averaged approach since, after normalisation (which corrects for field differences), the aggregation step should not differentiate between papers from different fields. Since the globalised approach gives more weight to ratios of papers that have higher expected citations counts, it introduces bias towards fields with high expected number of citations. The same argument was put forward by Lundberg (2007) since the globalised approach, in the case of the author level, does not normalise scores by individual paper scores, but on a higher aggregation level (in this case fields) where the average citation rate of an author is compared to the average citation rate of the fields in which the author publishes. However, the averaged approach uses the arithmetic mean and since citation distributions tend to be highly skewed, results may be influenced unfairly by a small set of highly cited papers. Similarly, recently published papers with only a few citations may be favoured unfairly since they are compared to expected citation counts that tend to be close to zero (Waltman et al., 2011a). However, the globalised approach also suffers from this problem (Opthof & Leydesdorff, 2010). Therefore, Lundberg (2007) argues for an item-oriented (averaged) normalisation based on a logarithmic *z*-score to overcome the typical skewness of citation distributions. He shows that with this approach, the impact of outliers decreases. Using this transformation, entity scores start to approach a normal distribution on research group levels (i.e., within department or university level).

Consistency is a property of metrics that postulates that the relative ranking of two entities should not change when both entities make the same progress in terms of papers and citations (Waltman et al., 2011a). The averaged approach is consistent whereas the globalised approach is not. However, Egghe and Rousseau (1996) argue that the globalised approach is more sound from a mathematical-statistical viewpoint.

Waltman, van Eck, van Leeuwen, Visser, and van Raan (2011b) show that the difference between the averaged and globalised approaches is very small at the country level and for large institutions but that the difference is somewhat larger for research group and journal levels. Larivière and Gingras (2011) find that the difference at any aggregation level is statistically significant and that it depends on the unit's number of papers. They show that the difference between these two approaches is greater for departments than for individual researchers. They argue that this is because the diversity of topics in which departments publish is generally greater than that of individuals.

2.3. Metrics

In addition to using citation counts (Citations) as impact scores, we also evaluate various other paper-level metrics that may be aggregated to the author level. We selected the specific metrics discussed below since each is based on a different ranking or normalisation paradigm. Therefore, each metric contains some unique feature which may be differently impacted by the averaged and globalised aggregation approaches.

We use a percentile approach (R6) in comparison to mean-normalised citation scores. In this approach papers are ranked according to their citation counts separately for each field and year. Papers are then assigned weights ([1, 2, 3, 4, 5, 6]) depending on which percentile intervals ([0–50, 50–75, 75–90, 90–95, 95–99, 99–100]) they belong to (Leydesdorff et al., 2011). Furthermore, we use the method suggested by Waltman and Schreiber (2013) to compute fair weights for papers across interval boundaries. The expected score for percentile metrics without weight classes is ideally always 50. This is not the case if the majority of papers have scores of 0 or if, as in our case, weight classes are used. Therefore, the expected score of a paper according to the R6 metric is the mean weight of papers in the same field and year.

We also use the relative citation ratio (RCR) metric since it uses the co-citation network of papers to normalise citation counts (Hutchins et al., 2016). More specifically, the actual score of a paper is its citation count normalised by its age. A paper's expected score is the average journal citation rate with a two-year citation windows of the journals at which it's co-cited papers are published.

Instead of only considering direct citations, we use two indirect metrics which consider the indirect impact of papers through reference chains. Most indirect metrics are recursively defined and take the entire structure of citation networks into account. The idea of recursively defining impact metrics originates from Pinski and Narin (1976). They applied it on academic citation networks to compute importance values for journals and to address the limitation that all citations are valued the same. The rationale of applying indirect metrics to citation networks is that citations from influential papers should count more than citations from unimportant papers.

The first indirect metric we consider is PageRank (Brin & Page, 1998) which is defined recursively and computes the probability of observing a paper when randomly traversing the citation graph. In contrast to the other metrics we consider, PageRank internally normalises the score of citations by the number of references of the citing paper. It is commonly known that PageRank is biased towards older papers in the paper citation graphs (Chen, Xie, Maslov, & Redner, 2007). We therefore normalise papers' PageRank scores over fields and publication years in the same way as done for citation counts.

The second indirect metric is the *Abramo* method which also takes the impact of citing papers into consideration (Giuffrida et al., 2018). However, differently to PageRank, it only considers two citation levels and transfers less of the citing paper's impact to the cited paper, by implementing the restriction that the gain through citations from high-impact papers should not be more than 1. Therefore, a paper with C citations has a score between C and $2C$. A score of $2C$ is achieved if all citing papers are themselves the highest cited papers among their respective fields and publication years. This metric is also field- and year-normalised. Comparing the Abramo method to mean-normalised citation counts may yield insight into how the inclusion of impact scores from indirect citations changes the results.

3. Methodology

3.1. Publication databases

We use two publication databases for the experiments described in this paper. The first is a 2015 version of the ACM Digital Library (ACM, Inc, 2014). It contains papers up to March 2015 that are published in periodicals and proceedings from the field of computer science. The ACM uses a categorisation scheme which is called the Computing Classification System (CCS) where each paper is associated with one or more concepts that are organised in a poly-hierarchical structure (ACM, Inc, 2017b). Each concept belongs to one or more parent concepts. Of all papers in the database, 703 802 have at least one associated concept (with an average of 4.49 concepts per paper). Since each concept is part of the hierarchical structure it can be collapsed such that each concept is associated with at least one top-level concept.

There are 13 top-level concepts in the CCS which we use as fields (see Table A.7 in A). We collapse the CSS concept structure such that papers are only associated with corresponding top-level concepts. A paper's field is then the top-level concept with which it is associated most frequently. This decision is based on the assumption that if a paper is associated with a top-level concept multiple times, it is closely related to that top-level concept and can therefore be categorised as such. For papers with equal frequencies of top-level concepts, we categorise them into each of the most occurring top-level concepts. We found that of all papers with concepts, 77% have a single most occurring top-level concept. All papers without concepts we classify into a separate field named 'none'. This decision is based on the fact that too many citations would be removed if all papers without concepts were removed. We remove papers that are not associated with a journal, conference series, or publication date. All authors associated with this subset of papers are extracted and assigned to one or more fields based on the number of papers they have published in a top-level concept. If an author has published the same number of papers in multiple fields, the author is assigned to each field. The resulting dataset comprises 1 737 687 papers and 861 550 authors.

For the Microsoft Academic Graph (MAG) database, we follow the same categorisation approach as with the ACM database. The MAG database includes a topic classification scheme with four hierarchical levels. However, we only use papers that are associated directly to one of the 19 top-level fields which can be interpreted as broad academic disciplines (see Table A.8 in A for the MAG field sizes), since lower-level fields in MAG are too noisy for citation analysis (Hug, Ochsner, & Brändle, 2017; Vaccario, Medo, Wider, & Mariani, 2017). The final MAG database comprises 13 829 901 papers and 14 314 921 authors.

3.2. Test data set

For the performance evaluation of the author impact metrics we use authors that are either ACM fellows or have won achievement or lifetime contribution awards (Dunaiski, 2018). We removed any duplicate entries in the test data set since some researchers are both ACM fellows and have won a lifetime contribution award (Dunaiski et al., 2018a). The resulting test data set sizes are 1 125 (ACM) and 1 220 (MAG). The ACM fellowship is the recognition of an individual's lasting impact on a field in computer science in terms of technical and leadership contributions, has influenced the direction of a field, and has to be evidenced by publications, awards or other publicly recognised artefacts of merit (ACM, Inc, 2017a). To become a fellowship candidate, researchers first have to be nominated by one of their peers who also submits endorsers. The nominator and the endorsers, collectively, have to be senior enough to make a credible case as to why a candidate's impact merits an ACM fellowship. For the authors that have won lifetime achievement awards, we considered awards that are handed out by conferences, learned societies, or special interest groups of academic disciplines. Generally, the nomination processes consist of peer nominations and final decisions are taken by dedicated award committees.

The assumption is that the entities in the test data exhibit some property that is not exclusively based on citations. Furthermore, we assume that their papers have had a lasting and influential impact on future papers. Therefore, we expect the test entities to have above average citation rates but also to have a latent property that is not encoded through pure citations. We use this test data to investigate whether certain metrics better identify the test entities and consequently their underlying property.

3.3. Performance evaluation

We aggregate paper scores to the author level and convert the resulting scores to fractional ranks where authors with tied scores are assigned their average rank (Dunaiski, Geldenhuys, & Visser, 2018b). For example, authors with the scores $\{10.5, 5.5, 5.5, 0.1\}$ would have the corresponding ranks of $\{1, 2.5, 2.5, 4\}$.

For the performance evaluation, we use the average ranking ratio to directly compare the metrics to each other in terms of ranking the test authors. This approach dampens the effect that outliers have on the results (Mariani, Medo, & Zhang, 2016). We use it to evaluate the metric's performance while considering different citation window sizes. In other words, which metrics better identify the well-established researchers t years after their first publication. The rank $r_i(m, t)$ of a test author i according to metric m is computed t years after their first publication. Similarly, the best rank by all considered metrics (m') is computed $\min_{m'}\{r_i(m', t)\}$. The ratio of these two values for each test author is averaged to yield the average ranking ratio $\bar{r}(m, t)$ of metric m for each citation window size t. For a citation window size of t, only citations that originate from papers published up to t years after author i's first year of publication are considered. The lower the ratio, the better a metric identifies the test authors. An optimal metric (with an average ranking ratio of 1 for all t) means that it always ranks the test authors higher than every other metric at each interval t.

3.4. Bias evaluation

The idea of the fairness test (Radicchi & Castellano, 2012) for metrics is to measure the deviation of the field distribution of the top p percent of papers to the field distribution of the overall sample of papers. For example, given a sample of 80 computer science and 20 mathematics papers, a fair metric would score the papers in such a way that the top 10% of papers

Table 3
The metrics' bias (d_M) for each variant on the ACM and MAG databases. As baselines, the first two rows show the bias of non-normalised citation counts and PageRank.

	d_M (ACM)			d_M (MAG)		
Metric	Total	Averaged	Globalised	Total	Averaged	Globalised
Citations (not normalised)	104.68	56.87[†]	–	450.04	202.61[†]	–
PageRank (not normalised)	102.37	45.57[†]	–	346.10	55.66[†]	–
R6	106.92	46.84	13.27	388.28	148.78	149.98
RCR	94.30	47.30	43.97	287.05	138.80	132.33
Abramo	101.85	26.00	26.82	394.27	160.14	161.94
Citations	102.81	25.26	24.88	384.23	156.72	157.74
PageRank	103.57	26.78	28.44	345.40	58.96	59.06

[†] Results are based on the average non-normalised paper score per author since expected paper scores are not defined for non-normalised metrics.

comprise 8 computer science and 2 mathematics papers. However, it is important to note that this fairness test may be biased itself if the category scheme used to normalise a metric is identical to the one used for measuring its fairness (Sirtes, 2012).

We use a method proposed by Vaccario et al. (2017) which uses the Mahalanobis distance (d_M) (Mahalanobis, 1936) to quantify the deviation between two distributions and is based on the assumption that a ranking is unbiased if its properties are consistent with that of an unbiased sampling process (Vaccario et al., 2017). We apply this approach to the author-level using the author categories as described in Section 3.1. Therefore, a percentage p from all authors N are randomly drawn without replacement. Based on this sample, the frequency that an associated field is observed is recorded in a vector $\boldsymbol{k}$. Let n be the number of authors constituting p% of all authors. Then the multivariate hypergeometric distribution gives the probability of observing such a vector $\boldsymbol{k}$

$$P(\boldsymbol{k}) = \frac{\prod_{f \in F} \binom{K_f}{k_f}}{\binom{N}{n}} \tag{1}$$

where F is the set of fields and K_f is the total number of authors in field f. An unbiased ranking would yield the expected number of authors for a field f as $\mu_f = n \cdot K_f/N$. Let $k_f^{(m)}$ be the number of top p% of authors in field f from an actual ranking metric m.

To quantify the deviation of the observed vector $\boldsymbol{k}^{(m)}$ from the expected vector $\boldsymbol{\mu}$, Vaccario et al. (2017) propose to simulate n_{sim} number of unbiased selection processes to obtain a set of ranking vectors distributed according to Equation (1) around the vector of expected values $\boldsymbol{\mu}$. For each vector in this distribution they compute the Mahalanobis distance to $\boldsymbol{\mu}$ to obtain a distribution of Mahalanobis distances from which one can compute confidence intervals. For every experiment in this paper, we simulate a number of unbiased sampling processes in which $p = 1\%$ of authors are sampled to create the statistical null model to which the metrics are compared. Authors are considered multiple times in the bias analyses if they belong to more than one field.

4. Results

We use the different paper-level metrics discussed in Section 2.3 and aggregate them to the author level by using the three aggregation approaches discussed in Section 2.1. Therefore, each metric has three variants: (1) the size-dependent (total) variant which is the sum of an author's paper scores, (2) the average of ratios (averaged) variant, and (3) the ratio of averages (globalised) variant.

We evaluate the author metrics on their field bias (Section 4.1) and ranking performance (Section 4.2). We compare the overall bias of the metrics and their different aggregation variants but focus on the difference between the two size-independent (averaged and globalised) variants while considering different citation window sizes. In terms of performance, we again focus on the difference between the averaged and globalised variants of each metric while considering different citation window sizes.

4.1. Bias

Table 3 shows the field bias of the metrics based on the ACM and MAG databases. We list the bias results of non-normalised PageRank and citation counts in the first two rows for comparison. Considering the size-dependent variants of the metrics (Total), the RCR metric has the least bias on both the ACM (94.30) and the MAG (287.05) databases. On the ACM database, all other metrics have relatively similar bias values. On the MAG database, however, standard PageRank (346.10) is substantially

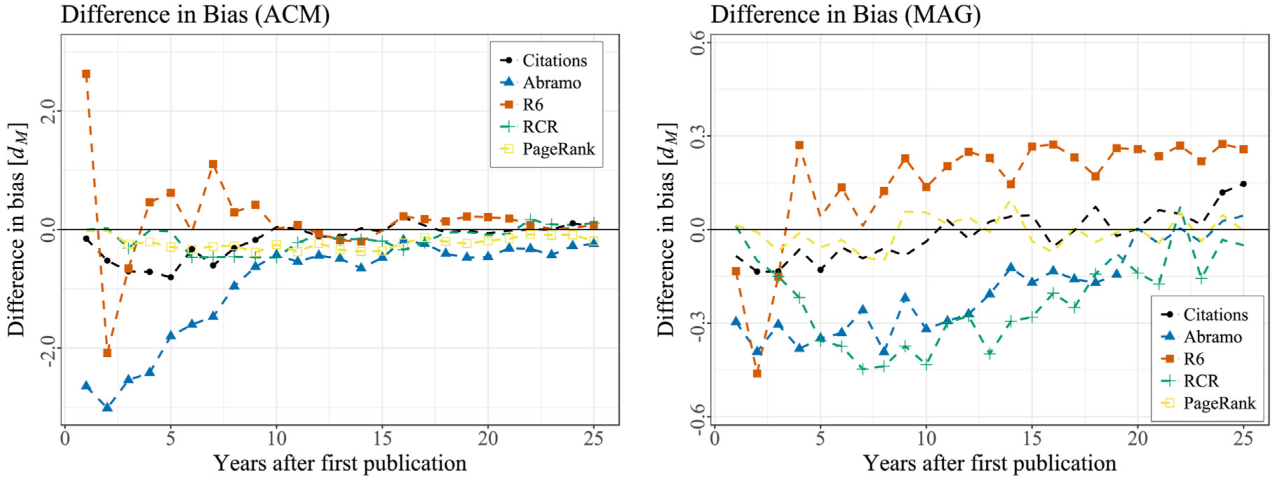


Fig. 1. The difference in bias between the averaged and globalised variants of the metrics plotted against the number of years since the authors' first publications. A negative (positive) value means that the averaged (globalised) approach has less bias. A marker on a curve denotes that the observed difference between a metric's variants is significant ($p < 0.05$) for that year.

less biased than citation counts (450.04). The bias of citation counts decreases when normalised for fields and years (384.23), but is still higher than PageRank's bias which remains about the same (345.40).

The size-independent variants (Averaged and Globalised) of the metrics have substantially less field bias compared to their size-dependent counterparts. The difference in bias between the averaged and globalised variants of the metrics is very small except for the R6 metric on the ACM database where the globalised variant has substantially less bias (13.27) compared to the averaged variant (46.84).

The plots in Fig. 1 show the differences in bias between the metrics' averaged and globalised variants plotted against the number of years since authors' first publication years. For each citation window size, we computed a statistical null model for author categories based on the papers they have published within the corresponding time period. Therefore, authors may swap categories if they start publishing predominantly in a different field as before. At each time interval we simulate 10 000 unbiased sampling processes in which 1% of authors are sampled to create the null models to which the metrics are compared.

In order to test for significance of the bias differences at each time interval, we create 1 000 bootstrap samples by randomly sampling 10% of the top 1% of ranked authors. We use Welch's t-test for unequal variances to test for significance and denote significant differences ($p < 0.05$) with markers on the curves in Fig. 1. For example, the left plot shows the mean difference between the averaged and globalised variants on the ACM database. For citation counts (round markers), the difference is only significant for the first 9 years, after which it is not (with the exception of a citation window size of 24). For PageRank, the difference is significant for all except the first two years.

On the ACM database, the averaged variants are less biased except for the R6 metric for which the bias values fluctuate. The difference in bias between the averaged and globalised variants is significant for most citation window sizes, except for citation counts which is only significant for the first 9 years. On the MAG database, the differences are not significant for PageRank and citation counts for most citation window sizes. The globalised variant of R6 shows less bias except for the first three years. Lastly, the averaged variants of RCR and the Abramo method have less bias on both databases. In general, the differences between the metrics' variants are similar on both databases. PageRank and citation counts have the smallest differences between the variants. The variants of the Abramo method seem to be more dependent on citation window sizes compared to the other metrics since the differences are relatively large for small citation windows which decrease with larger citation windows.

4.2. Performance

Fig. 2 shows the average ranking ratio of the metric's size-dependent variants (total) on the ACM and MAG databases. The trends on both databases are relatively similar. RCR performs the best in identifying the well-established researchers early in their careers. However, the RCR metric performs the worst after 9 years on the ACM database. On the MAG database, however, only PageRank performs better than RCR for larger citation windows ($t > 12$). PageRank performs the best in identifying the well-established researchers in later stages of their careers which is after 15 (ACM) and 12 (MAG) years.

For the first 10–12 years, the R6 metric performs relatively similar to the Abramo method, after which the Abramo method performs consistently better. Since the Abramo method defines a paper's score as its normalised citation count plus an additional score based on indirect impact, it is interesting to compare its results to citation counts. On the ACM database, citation counts perform better for smaller citation windows. The Abramo method surpasses citation counts after 14 years and shows an observable performance difference. On the MAG database, however, both methods perform similarly for larger citation windows.

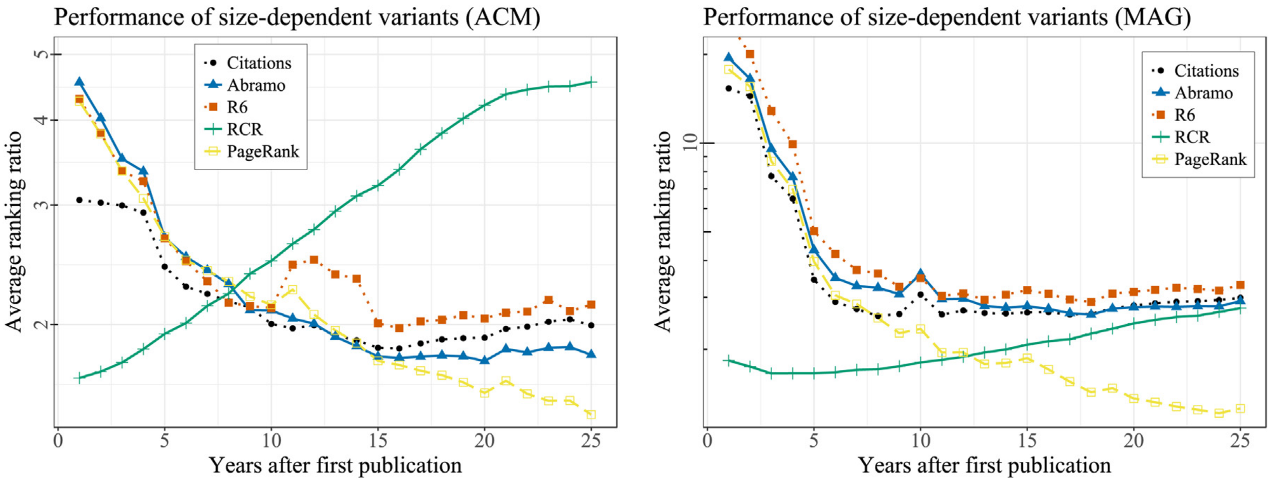


Fig. 2. The average ranking ratio of the metrics's total variants plotted against the time since the test authors' first year of publication on the ACM (left) and MAG (right) databases.

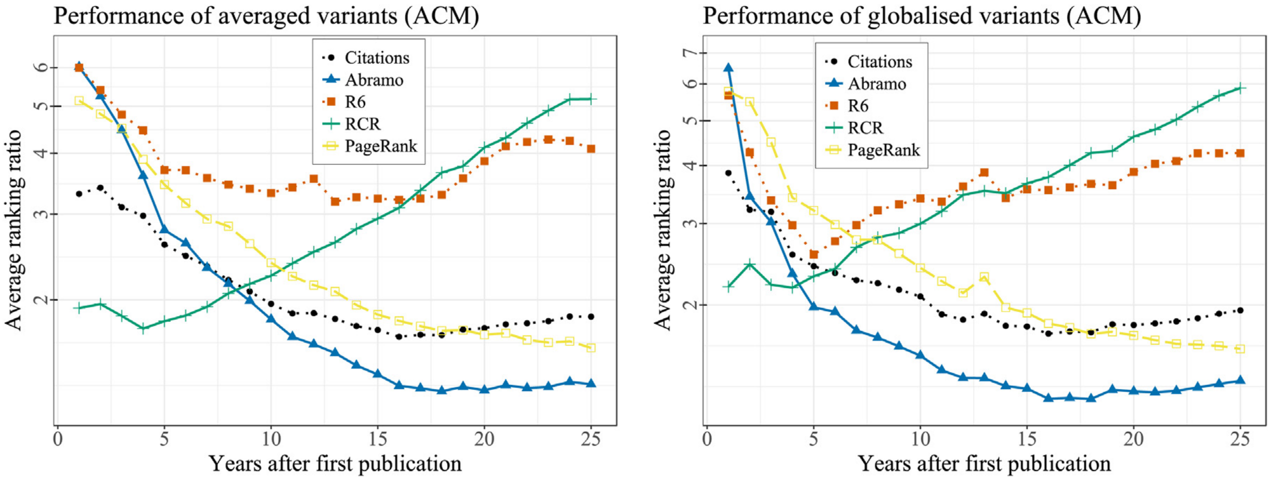


Fig. 3. The average ranking ratio of the metrics' averaged (left) and globalised (right) variants plotted against the time since the test authors' first publication years on the ACM database.

The plots in Fig. 3 show the performance of the metrics' averaged (left) and globalised (right) variants on the ACM database. The size-independent variants' relative performances are similar to their total counterparts, however, a few differences exist. PageRank no longer performs better than the Abramo method for larger citation windows. Furthermore, the RCR metric only performs the worst for citation windows larger than 16 years compared to 9 years for the total variants. The relative performances of the metrics when comparing the averaged variants to the globalised variants are very similar. However, when the globalised variants are compared, then the Abramo method performs better than citation counts for all citation window sizes except for the first two year. Furthermore, the performance of the percentile approach (R6) is better than PageRank for the first six years.

The plots in Fig. 4 for the MAG database are analogous to the plots in Fig. 3 for the ACM database. When comparing these sets of plots a few differences exist. For both variants, PageRank performs better on the MAG database compared to citation counts and the Abramo method. Furthermore, the R6 metric performs the worst for all citation windows.

As on the ACM database, the relative performances of the metrics are very similar when comparing the averaged variants to the globalised variants. The biggest difference is again between citation counts and the Abramo method. The averaged variant of citation counts performs better than the Abramo method for the first 11 years after which it perform slightly worse. However, when comparing the globalised variants, citation counts only perform better for the first four years.

In Fig. 5 we show the performance differences of the metrics when the averaged variants are directly compared to the globalised variants. The plots show the difference in average ranks between the globalised and averaged variants for the ACM (left) and MAG (right) databases. Furthermore, to analyse whether the difference in the average rank at each time interval is significant, we use a bootstrapped approach to obtain means and variances of performance differences. We compute the performance differences of 1 000 bootstrap samples by randomly sampling with replacement 10% of the test entities. We use the observed performance differences to test for significance using the Welch's *t*-test for unequal variances. Significant differences ($p < 0.05$) are denoted by the markers in the plots.

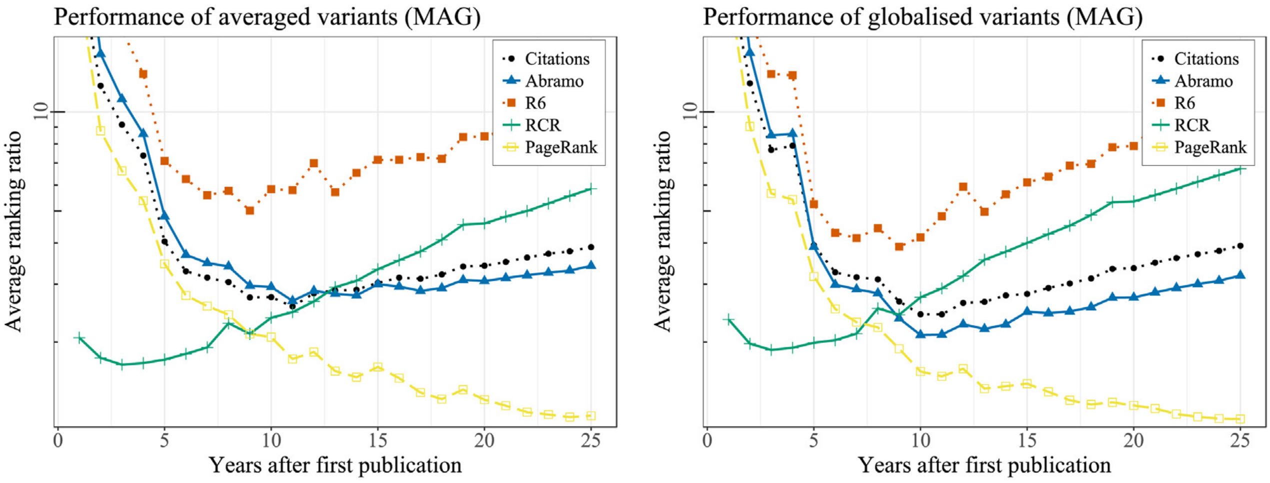


Fig. 4. The average ranking ratio of the metrics' averaged (left) and globalised (right) variants plotted against the time since the test authors' first publication year on the MAG database.

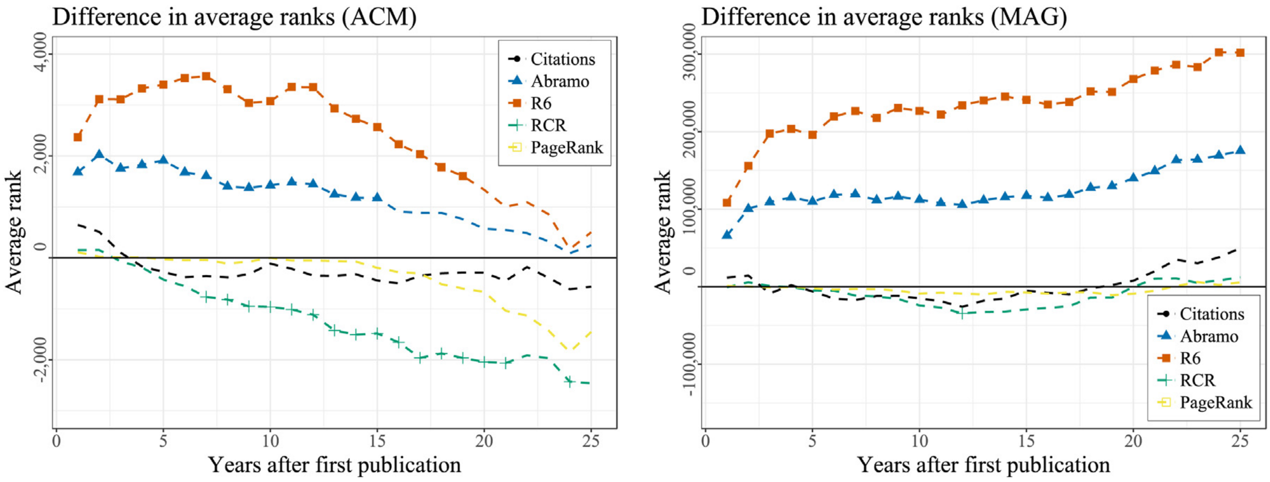


Fig. 5. The difference in performance between the averaged and globalised variants of the metrics plotted against the number of years since the test authors' first publications. Negative (positive) values mean that the averaged (globalised) variant performs better. Markers on a curve denote a significant ($p < 0.05$) difference between a metric's variants.

Table 4
Summary of the bias and performance results. The best performing (average ranking ratio) and least biased metrics are listed for different citation window sizes *t* for the ACM and MAG databases.

		ACM			MAG		
	t	Total	Averaged	Globalised	Total	Averaged	Globalised
Performance	2	RCR	RCR	RCR	RCR	RCR	RCR
	5	RCR	RCR	Abramo	RCR	RCR	RCR
	10	Citations	Abramo	Abramo	PageRank	PageRank	PageRank
	25	PageRank	Abramo	Abramo	PageRank	PageRank	PageRank
Bias	2	RCR	PageRank	PageRank	RCR	PageRank	PageRank
	5	RCR	R6	R6	RCR	PageRank	PageRank
	10	RCR	R6	R6	RCR	RCR	RCR
	25	RCR	Abramo	Citations	RCR	PageRank	PageRank,R6

For PageRank and citation counts the performance differences between the variants is not significant for any citation window size on both databases. The globalised variants of the R6 metric and the Abramo method perform better. On the ACM database they perform significantly better for the first 15 (Abramo) and 19 (R6) years, after which the differences are no longer significant and tend towards 0. On the MAG database, the globalised variants also perform significantly better with a slight upwards trend in favour of the globalised variants. On the ACM database, the performance differences between RCR's variants are not significant for the first six years, after which the averaged variant performs better. On the MAG database, however, the differences are not significant except for a citation window size of 12 years.

Table 4 briefly summarises the results when comparing the different metrics directly to each other. For different citation window sizes t, the table lists the best performing (average ranking ratio) and least biased (d_M) metric on the ACM and MAG databases.

5. Discussion

The Abramo method. As mentioned before, comparing the results of the Abramo method to citation counts may yield some insight into how the indirect impact can influence the performance and bias of author scores. The Abramo method defines a paper's score as its citation count C plus the score obtained from indirect citations which ranges between 0 and C. Therefore, the actual score of a paper ranges between C and $2C$, where $2C$ is achieved if all citing papers are themselves the most cited papers in their respective year and field categories. The expected score of a paper is the mean actual score of papers from the same field and published in the same year. The additional indirect impact score of the current paper is field and year independent. It only depends on the citing papers' citation counts and their positions on the respective field and year citation distributions. Therefore, papers' scores including their gains through indirect citations are compared to the expected score of a field which includes the field's average gain through indirect citations. Consequently, papers with a large gain through indirect citations also have a comparatively larger value when computing the ratio of actual and expected scores. Similarly, papers with little or no gain through indirect citations are compared to a higher expected score than the average citation counts of their fields. Therefore, the differences between high and low scoring papers are increased compared to standard citation counts. Furthermore, the Abramo method creates higher expected ratios for fields that contain, on average, papers with higher indirect impact scores.

On the paper level, Giuffrida et al. (2018) show that the correlation between citation counts and scores of the Abramo method is very high but that outliers are relatively frequent. In terms of overall bias, we found that the two metrics are very similar with citation counts showing slightly less field bias (Table 3). Globalised variants favour ratios with high expected scores (Lundberg, 2007). However, on the author level, we assume that the majority of authors publish in a very small number of subject categories (Larivière & Gingras, 2011). Therefore, the impact of higher expected scores of the Abramo method is somewhat reduced by normalising with the gain-included expected scores. However, we hypothesise that the difference between the averaged and globalised variant is much larger at higher aggregation levels where research units publish in more diverse fields or when databases are used with finer-grained field delineations. Furthermore, we believe that the field bias of the globalised variant will increase compared to standard citation counts.

In terms of performance of the size-dependent variants (total), we found that citation counts perform better for citation windows up to 12–15 years. For the averaged variants, the Abramo method only performs better when $t > 7$ (ACM) and $t > 12$ (MAG). Interestingly, the Abramo method performs better than citation counts when the globalised variants are compared. Furthermore, the differences between the two variants is only significant for the Abramo method. Further investigation is required to understand whether the better performance of the globalised variant of the Abramo method is due to the test entities coming from fields with higher expected scores or whether this is due to the test authors' papers receiving above average gains through indirect citation impacts. In summary, the differences between the two variants is negligible for citation counts. However, for the Abramo method, the averaged variant shows less bias while the globalised variant identifies the well-established researchers better than citation counts.

The RCR metric. The expected score of a paper computed by the RCR metric is the average citation rate of the journals in which the paper's co-cited papers are published. Since the globalised variants give more weight to ratios with high expected scores, ratios from papers contribute more towards authors' scores if their co-cited papers are published in highly cited journals. The rationale for the RCR metric of using co-cited papers as reference sets is that they are topically similar since they were cited together to produce new work. This may be justified for creating reference sets on the paper level even though many co-cited papers are from different topics (Janssens et al., 2017). On the one hand, it may be argued that on the author level the dependence on the journal citation rate at which the reference papers were published should not influence the author's impact score, especially if they are from unrelated disciplines. On the other hand, one may argue that a paper that performs well with respect to a cohort of papers published at prestigious journals should in fact count more towards an author's impact score.

PageRank. The overall bias of non-normalised PageRank on the MAG database is less than that of non-normalised citation counts but on the ACM database the field bias is about the same. This indicates that, on a multidisciplinary database, PageRank manages to normalise scores through its internal (citing-side) normalisation step in which the impact of a citation is normalised by the number of references in a paper's reference list. Interestingly, PageRank's bias does not decrease substantially when it is normalised over fields and time.

The R6 metric. It should be noted that a percentile approach without weights, where the median paper always scores 50%, would result in no difference between the averaged and globalised variants. However, we found that with the weights of the R6 metric, the globalised variant performs better in terms of both field bias and ranking performance.

6. Threats to validity and future work

In Section 4.1 we computed the bias of metrics for differently sized citation windows t. For this computation we created the null models based on all authors that have received a score at time $t_1 + t$, where t_1 is the year in which an author first published a paper. The score of an author is therefore based on a citation graph that contains all papers and citations up to the year $t_1 + t$. Since t_1 varies among authors, the author scores at time t are based on citation graphs with different time ranges $[t_0, t]$, where t_0 is the overall minimum year which we set to 1800 (MAG) and 1950 (ACM). This allows us to only compute $|2017 - t_0|$ copies of scores for each metric with citation graph year ranges of $[t_0, t_i]\ \forall\ t_i \in t_0, \ldots, 2017$.

Since we compare scores from different citation graphs at the time interval t, we have to convert all author scores into a single list of new ranks. Therefore, authors are compared that have the same career lengths t but from different time periods and different citation graph sizes. However, this does not impact the results of direct metrics since we normalise all metrics by year and each paper starts with a citation count of 0. The same is true for the Abramo method. However, this decision might have a small impact on the PageRank results. PageRank is computed over a citation graph with n nodes where n edges with weights $1/n$ are added from dangling nodes to each paper. Therefore, the size of the underlying citation graph does impact the final scores of the papers to some degree. An alternative approach is to use individual citation graphs that only include papers and citations from the year range $[t_i, t_i + t_w]\ \ \forall\ \ t_i \in t_0, \ldots, 2017\ \ \forall\ \ t_w \in 1, \ldots, 25$. However, this approach also compares PageRank scores based on differently sized citation graphs since the size of the academic corpus grows at an increasing rate. Furthermore, it is computationally very demanding since, in the case of the MAG database, PageRank would have to be computed on 5125 different citation graphs. In addition, all metrics would have to be compared to 5125 different null models.

It should be mentioned that the two evaluation criteria (bias and ranking performance) used in this paper are not exhaustive evaluation standards. The choice between the averaged and globalised variants might change if different evaluation criteria are applied. For example, one may argue that a violation of the consistency property (Waltman et al., 2011a) is unacceptable and that the globalised variants should therefore be rejected. For future investigations, additional evaluation criteria and alternative methods should be considered. For example, the methodology proposed by Crespo, Li, and Ruiz-Castillo (2013) measures the citation inequality between fields that stems from varying citation practices by comparing multiple quantiles of citation distributions. Applying this methodology might yield further insight into the bias differences between the globalised and averaged approaches. Furthermore, by using finer-grained or algorithmically defined subject categories (Crespo, Herranz, Li, & Ruiz-Castillo, 2014), further characteristics of the metrics and their aggregated variants may be identified.

We would like to highlight that the performance results should not be interpreted outside the scope of the test data. Since the performance results of the metrics are based on test data comprising well-established researchers the results might be different for researchers in other stages of their careers. To generalise the results, we recommend that the analyses presented in this paper are extended with additional test data comprising different attributes. Furthermore, this type of study may also be applied to higher aggregation levels (e.g., journal or institution) if appropriate external test data is available.

7. Conclusion

In this paper, we compared the averaged (average of ratios) and the globalised (ratio of averages) aggregation approaches on the author level. We used different time- and field-normalised paper-level metrics which we, once aggregated to the author level, evaluated in terms of field bias and performance. We evaluated performance based on how well the metrics rank the authors of a test data set comprising well-established researchers who have received accolades for their impactful and long-lasting contributions in their respective fields.

We found no significant difference between the averaged and globalised approach for citation counts in terms of bias and performance. Similarly, we found little difference between the two approaches with paper scores based on PageRank. However, for the percentile metric (R6) the globalised approach is the better choice. The Abramo method shows less bias with the averaged approach but identifies the well-established researchers better with the globalised approach. Lastly, we found that the RCR metric's averaged variant exhibits less bias but the performance difference between the variants is insignificant on the multi-disciplinary database.

In terms of overall bias, the R6 metric has the least bias on the computer science database. However, on the multi-disciplinary database, PageRank has the least bias which indicates that its citing-side normalisation step successfully reduces field bias. This result gives further support for citing-side normalised approaches. We recommend that further analyses of this kind are conducted to compare different metrics that incorporate citing-side normalisations. In terms of performance, the RCR metric identifies the well-established researchers best for small citation windows. For larger citation windows, the metrics that incorporate indirect impacts through reference chains outperform metrics that only consider direct citations. This indicates that valuable information is encoded in citation networks that should be leveraged to better identify well-established researchers which have had continued high impact in their fields.

Author contributions

Marcel Dunaiski: Conceived and designed the analysis, Collected the data, Contributed data or analysis tools, Performed the analysis, Wrote the paper.

Jaco Geldenhuys: Other contribution.

Appendix A. Supplementary information

Table A.5
The field bias of the metrics' variants at different citation windows sizes t and the average ranks of the test entities based on the ACM database. The last two row give the number of authors and fields in the corresponding null models. The bias values are the normalised Mahalanobis distances $(d_M/\sqrt{F*(n-1)})$ for distributions of different sizes.

ACM		Bias				Avg. Rank			
Metric	Variant	$t=2$	$t=5$	$t=10$	$t=25$	$t=2$	$t=5$	$t=10$	$t=25$
Citations	Total	0.201	0.255	0.291	0.306	9 716	5 804	3 964	5 812
R6	Total	0.240	0.276	0.300	0.309	11 048	6 144	3 439	4 204
RCR	Total	0.141	0.197	0.226	0.280	9 272	5 695	4 227	8 385
Abramo	Total	0.207	0.261	0.291	0.300	10 500	6 025	3 780	4 970
PageRank	Total	0.238	0.272	0.299	0.303	8 796	5 509	3 670	4 571
Citations	Averaged	0.068	0.082	0.082	0.075	10 963	10 339	13 734	45 208
Citations	Globalised	0.074	0.089	0.082	0.077	10 417	10 642	13 862	45 833
R6	Averaged	0.090	0.042	0.058	0.099	14 490	14 358	18 354	53 350
R6	Globalised	0.111	0.033	0.058	0.097	11 343	11 013	15 286	52 936
RCR	Averaged	0.063	0.066	0.087	0.100	10 370	10 353	16 780	71 584
RCR	Globalised	0.064	0.067	0.093	0.093	10 234	10 908	17 955	74 061
Abramo	Averaged	0.085	0.092	0.083	0.072	12 520	11 159	12 742	35 223
Abramo	Globalised	0.114	0.111	0.090	0.085	10 448	9 265	11 305	34 976
PageRank	Averaged	0.051	0.079	0.073	0.091	10 742	12 679	17 971	60 189
PageRank	Globalised	0.052	0.082	0.078	0.102	10 716	12 758	18 054	61 569
Size of null model (n):		7 733	6 631	3 638	620				
Number of fields (F):		14	14	14	14				

Table A.6
The field bias of the metrics' variants at different citation windows sizes t and the average ranks of the test entities based on the MAG database. The last two row give the number of authors and fields in the corresponding null models. The bias values are the normalised Mahalanobis distances $(d_M/\sqrt{F*(n-1)})$ for distributions of different sizes.

MAG		Bias				Avg. Rank			
Metric	Variant	$t=2$	$t=5$	$t=10$	$t=25$	$t=2$	$t=5$	$t=10$	$t=25$
Citations	Total	0.099	0.137	0.165	0.193	620 084	470 523	356 711	388 149
R6	Total	0.136	0.158	0.163	0.180	777 635	516 045	342 692	323 564
RCR	Total	0.049	0.065	0.094	0.135	620 493	440 213	338 443	393 650
Abramo	Total	0.106	0.139	0.171	0.199	705 576	504 298	364 682	371 527
PageRank	Total	0.108	0.120	0.127	0.152	549 403	413 184	273 621	240 885
Citations	Averaged	0.081	0.083	0.077	0.055	639 231	624 363	735 733	1 648 303
Citations	Globalised	0.081	0.083	0.077	0.053	624 204	629 002	747 945	1 601 852
R6	Averaged	0.078	0.068	0.061	0.053	912 304	870 713	1 028 000	2 172 260
R6	Globalised	0.079	0.068	0.060	0.049	754 747	674 370	798 240	1 872 812
RCR	Averaged	0.079	0.056	0.050	0.064	651 247	594 014	721 166	1 719 726
RCR	Globalised	0.079	0.058	0.052	0.063	638 945	592 647	737 253	1 695 824
Abramo	Averaged	0.087	0.086	0.080	0.058	775 427	718 422	795 154	1 581 474
Abramo	Globalised	0.089	0.088	0.082	0.058	674 276	608 604	679 632	1 409 106
PageRank	Averaged	0.039	0.048	0.056	0.047	518 163	554 926	602 082	1 015 376
PageRank	Globalised	0.039	0.048	0.056	0.048	517 137	554 854	605 752	1 013 065
Size of null model (n):		292 869	244 782	158 421	41 943				
Number of fields (F):		19	19	19	19				

Table A.7
The top-level concepts of the ACM database with their number of papers, authors, and test authors.

Concept	Papers	Avg. Citations	Authors	Avg. Citations	Test Authors	Avg. Citations
Applied computing	98 794	2.85	62 713	2.17	17	515.65
Computer syst. org.	31 583	6.30	23 098	4.15	39	1 411.46
Computing methods	156 981	6.56	122 815	5.30	196	1 483.41
General and refs.	33 237	2.42	7 739	2.23	3	232.33
Hardware	50 400	4.35	36 101	4.61	72	1 686.18
Human-cent. comp.	43 369	6.30	35 698	4.52	48	1 718.50
Information systems	90 588	7.30	63 114	6.33	177	2 191.94

Table A.7 (*Continued*)

Concept	Papers	Avg. Citations	Authors	Avg. Citations	Test Authors	Avg. Citations
Mathematics of comp.	127 090	4.16	63 458	3.72	71	824.00
Networks	46 344	5.98	36 830	5.42	95	1 918.36
Security and privacy	14 230	7.05	10 219	7.72	21	1 635.19
Social and prof. topics	54 319	3.36	26 683	3.00	40	360.40
Software and its eng.	92 507	6.99	57 414	6.71	241	1 572.66
Theory of comp.	84 281	7.35	39 867	7.48	170	1 693.12
None	1 033 885	2.45	419 152	0.91	4	20.50
Total Papers:	1 957 608	Total Authors:	1 004 901	Test Authors:	1 194	
Unique Papers:	1 737 687	Unique:	861 550	Unique:	1 125	

Table A.8
The top-level fields of the MAG database with their number of papers, authors, and test authors.

Field	Papers	Avg. Citations	Authors	Avg. Citations	Test Authors	Avg. Citations
Art	137 526	10.54	99 802	8.70	1	16.00
Biology	4 935 116	19.05	5 081 799	75.84	16	2 459.88
Business	423 171	10.90	317 595	10.54	1	80.00
Chemistry	4 975 005	14.98	3 953 081	36.70	35	1,233.60
Computer Science	3 002 769	13.56	2 146 175	20.46	979	1,181.08
Economics	1 636 621	12.27	1 251 746	15.79	11	691.55
Engineering	2 122 686	11.54	1 492 635	13.30	33	216.24
Environ. Science	240 439	26.42	155 077	23.63	0	–
Geography	220 697	14.31	141 392	12.36	0	–
Geology	1 448 707	15.90	996 747	30.50	5	618.80
History	277 970	12.65	210 754	11.78	1	7.00
Materials Science	1 563 279	13.35	981 126	20.08	3	190.00
Mathematics	3 533 736	12.20	2 256 061	18.51	158	954.51
Medicine	4 118 277	16.87	4 576 381	46.71	4	11.25
Philosophy	565 568	11.48	363 781	10.72	7	113.86
Physics	5 172 391	12.08	3 679 888	35.77	53	1,076.42
Political Science	80 126	7.29	58 740	7.65	4	42.75
Psychology	2 323 402	16.76	2 000 062	34.88	13	1,539.54
Sociology	1 303 427	12.06	993 677	13.73	14	197.14
Total Papers:	38 080 913	Total Authors:	30 756 519	Test Authors:	1 338	
Unique Papers:	13 829 901	Unique Authors:	14 314 921	Unique:	1 220	

Appendix B. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.joi.2019.01.006.

References

ACM, Inc. (2014). *ACM Digital Library.* , http://dl.acm.org/. [Online; accessed 12-Sep-2018]

ACM, Inc. (2017a]). *ACM Fellows.* , https://awards.acm.org/fellows/nominations. [Online; accessed 05-Sep-2017].

ACM, Inc. (2017b]). *The 2012 ACM Computing Classification System.* , http://www.acm.org/publications/class-2012-intro. [Online; accessed 12-Sep-2018].

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom - a test of Zitt's hypothesis? *Scientometrics*, *75*(1), 81–95.

Bornmann, L. (2017). Measuring impact in research evaluations: a thorough discussion of methods for, effects of and problems with impact measurements. *Higher Education*, *73*(5), 775–787.

Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits? *Journal of Informetrics*, *7*(1), 158–165.

Bornmann, L., & Marx, W. (2015). Methods for the generation of normalized citation impact scores in bibliometrics: Which method best reflects the judgements of experts? *Journal of Informetrics*, *9*(2), 408–418.

Bornmann, L., & Marx, W. (2018). Critical rationalism and the search for standard (field-normalized) indicators in bibliometrics? *Journal of Informetrics*, *12*(3), 598–604.

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.

Colliander, C. (2015). A novel approach to citation normalization: A similarity-based method for creating reference sets? *Journal of the Association for Information Science and Technology*, *66*(3), 489–500.

Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments? *Journal of Informetrics*, *5*(1), 101–113.

Crespo, J. A., Herranz, N., Li, Y., & Ruiz-Castillo, J. (2014). The effect on citation inequality of differences in citation practices at the web of science subject category level? *Journal of the Association for Information Science and Technology*, *65*(6), 1244–1256.

Crespo, J. A., Li, Y., & Ruiz-Castillo, J. (2013). The measurement of the effect on citation inequality of differences in citation practices across scientific fields? *PLOS ONE*, *8*(3), 1–9.

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks? *Journal of the American Society for Information Science and Technology*, *60*(11), 2229–2243.

Dunaiski, M. (2018). *Data for: Author ranking evaluation at scale. Mendeley data, v1*. https://doi.org/10.17632/5tzchw6r6d.1.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018a]). Author ranking evaluation at scale? *Journal of Informetrics*, *12*(3), 679–702.

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018b]). How to evaluate rankings of academic entities using test data? *Journal of Informetrics*, *12*(3), 631–655.

Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards? *Journal of Informetrics*, *10*(2), 392–407.

Egghe, L., & Rousseau, R. (1996). Averaging and globalising quotients of informetric and scientometric data? *Journal of Information Science*, *22*(3), 165–170.

Fiala, D. (2012). Time-aware PageRank for bibliographic networks? *Journal of Informetrics*, *6*(3), 370–388.

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), 334–348.

Fiala, D., & Tutoky, G. (2017). PageRank-based prediction of award-winning researchers and the impact of citations? *Journal of Informetrics*, *11*(4), 1044–1068.

Gao, C., Wang, Z., Li, X., Zhang, Z., & Zeng, W. (2016). PR-Index: Using the h-Index and PageRank for Determining True Impact. *PLOS ONE*, *11*(9), e0161755.

Giuffrida, C., Abramo, G., & D'Angelo, C. A. (2018). *Do all citations value the same? Valuing citations by the value of the citing items.* , https://arxiv.org/abs/1809.06088.

Hug, S. E., Ochsner, M., & Brändle, M. P. (2017). Citation analysis with microsoft academic? *Scientometrics*, *111*(1), 371–378.

Hutchins, B. I., Yuan, X., Anderson, J. M., & Santangelo, G. M. (2016). Relative Citation Ratio (RCR): A New Metric That Uses Citation Rates to Measure Influence at the Article Level. *PLoS Biology*, *14*(9), e1002541.

Janssens, A., Goodman, M., Powell, K., & Gwinn, M. (2017). A critical evaluation of the algorithm behind the Relative Citation Ratio (RCR). *PLoS Biology*, *15*(10), e2002536.

Larivière, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, *5*(3), 392–399.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the Tables on Citation Analysis One More Time: Principles for Comparing Sets of Documents? *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community? *Information Processing & Management*, *41*(6), 1462–1480.

Lundberg, J. (2007). Lifting the crown-citation z-score? *Journal of Informetrics*, *1*(2), 145–154.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 49–55.

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality? *Journal of Informetrics*, *10*(4), 1207–1223.

Microsoft. (2017). *Microsoft academic graph.* , https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/. [Online; accessed 15-Aug-2017].

Moed, H. F. (2010). CWTS crown indicator measures citation impact of a research group's publication oeuvre? *Journal of Informetrics*, *4*(3), 436–438.

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.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("leiden") evaluations of research performance? *Journal of Informetrics*, *4*(3), 423–430.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, *12*(5), 297–312.

Radicchi, F., & Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts? *Journal of Informetrics*, *6*(1), 121–130.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: toward an objective measure of scientific impact? *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17268–17272.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure? *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science? *Journal of Informetrics*, *9*(1), 102–117.

Silva, F. N., Rodrigues, F. A., Oliveira, O. N., da, F., & Costa, L. (2013). Quantifying the interdisciplinarity of scientific journals and fields? *Journal of Informetrics*, *7*(2), 469–477.

Sirtes, D. (2012). Finding the easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair? *Journal of Informetrics*, *6*(3), 448–450.

Smolinsky, L. (2016). Expected number of citations and the crown indicator? *Journal of Informetrics*, *10*(1), 43–47.

Vaccario, G., Medo, M., Wider, N., & Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network? *Journal of Informetrics*, *11*(3), 766–782.

van Leeuwen, T. N., & Calero Medina, C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics? *Research Evaluation*, *21*(1), 61–70.

Vinkler, P. (2012). The case of scientometricians with the "absolute relative" impact indicator? *Journal of Informetrics*, *6*(2), 254–264.

Waltman, L. (2015). *NIH's new citation metric: A step forward in quantifying scientific impact?* , https://www.cwts.nl/blog?article=n-q2u294. [Online; accessed 10-Sep-2018].

Waltman, L. (2016). A review of the literature on citation impact indicators? *Journal of Informetrics*, *10*(2), 365–391.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators? *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379.

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science? *Journal of the American Society for Information Science and Technology*, *63*(12), 2378–2392.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. (2011). Towards a new crown indicator: Some theoretical considerations? *Journal of Informetrics*, *5*(1), 37–47.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: an empirical analysis? *Scientometrics*, *87*(3), 467–481.

Waltman, L., Yan, E., & van Eck, N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics*, *89*(1), 301.

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.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation? *Scientometrics*, *63*(2), 373–401.

Journal of Informetrics 12 (2018) 679–702

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Author ranking evaluation at scale

Marcel Dunaiski[a,b,*], Jaco Geldenhuys[b], Willem Visser[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 9 October 2017
Received in revised form 1 June 2018
Accepted 10 June 2018

Keywords:
Author indicators
Author ranking
Citation analysis
PageRank

ABSTRACT

We evaluate author impact indicators and ranking algorithms on two publication databases using large test data sets of well-established researchers. The test data consists of (1) ACM fellowship and (2) various life-time achievement awards. We also evaluate different approaches of dividing credit of papers among co-authors and analyse the impact of self-citations. Furthermore, we evaluate different graph normalisation approaches for when PageRank is computed on author citation graphs.

We find that PageRank outperforms citation counts in identifying well-established researchers. This holds true when PageRank is computed on author citation graphs but also when PageRank is computed on paper graphs and paper scores are divided among co-authors. In general, the best results are obtained when co-authors receive an equal share of a paper's score, independent of which impact indicator is used to compute paper scores. The results also show that removing author self-citations improves the results of most ranking metrics. Lastly, we find that it is more important to personalise the PageRank algorithm appropriately on the paper level than deciding whether to include or exclude self-citations. However, on the author level, we find that author graph normalisation is more important than personalisation.

1. Introduction

To empirically answer questions about bibliometric indicators, a representative publication database and appropriate evaluation data (or test data) are required. One problem of evaluating indicators and algorithms that measure academic quality or impact is the difficulty of obtaining appropriate test data. This problem is compounded by the subjectivity of what is considered quality or impact, and generally requires human judgment. Due to this drawback, correlation analyses are often performed which are problematic on their own (Thelwall, 2016) and only provide comparatives to some baselines, usually citation counts used as a proxy for quality.

Another option that is often employed is the use of relatively small test data sets that are based on some external knowledge. The assumption is that the author entities in the test data exhibit some property (e.g., are highly influential or well-established) that is not exclusively based on citations. Therefore, these test data sets are often used to evaluate the functionality of the ranking algorithms to identify the comprising entities and consequently their shared property. Examples of such applications are: evaluating author ranking algorithms in identifying well-established researchers using test data that comprises researchers that have received fellowship status at learned societies, have won life-time contribution awards,

[*] Corresponding author at: Media Lab, Stellenbosch University, 7602 Matieland, South Africa.
E-mail address: marcel@ml.sun.ac.za (M. Dunaiski).

https://doi.org/10.1016/j.joi.2018.06.004
1751-1577/© 2018 Elsevier Ltd. All rights reserved.

or are frequently board members of prestigious journals (Dunaiski, Visser, & Geldenhuys, 2016; Fiala, Šubelj, Žitnik and Bajec, 2015; Fiala, Rousselot, & Ježek, 2008; Fiala & Tutoky, 2017; Gao, Wang, Li, Zhang, & Zeng, 2016; Nykl, Ježek, Fiala, & Dostal, 2014); evaluating the performance of paper-level ranking algorithms in finding impactful papers using test data that comprises best paper awards or high-impact paper awards (Dunaiski et al., 2016; Dunaiski & Visser, 2012; Mariani, Medo, & Zhang, 2016; Sidiropoulos & Manolopoulos, 2005); and to showcase the applicability of newly proposed indicators (Gao et al., 2016). Very rarely, direct peer-reviewed opinions are used to evaluate metrics (i.e., Abramo & D'Angelo, 2015) since this type of information is often not readily available.

Nykl et al. (2014) analyse various PageRank approaches and the effects that author graph normalisations and self-citations have on the ranking of authors. As evaluation data they use 54 authors that have won one of two prestigious computer science awards, a set of 576 researchers that have received fellowships of the Association for Computing Machinery (ACM) (ACM, Inc., 2017b), and a list of 280 highly cited researchers. The data they use for the experiments is a subset of the Web Of Science (Clarivate Analytics, 2017) publication data consisting of 149 347 papers published in 386 computer science journals between 1996 and 2005.

Later, Nykl, Campr, and Ježek (2015) extend this research to include different schemes to answer the question of how the credit of a paper should be shared among its co-authors. They again use the ACM fellows as evaluation data and two different lists of author names in the computer science fields of "artificial intelligence" and "hardware" with 354 and 158 authors, respectively. These lists comprise authors that have won contribution awards, but also authors that have written papers that have won best paper awards or influential paper awards, which are usually handed out about 10 years after initial publication for their outstanding impact in their fields.

In this paper, we reproduce and extend the above mentioned work by Nykl et al. (2014) and Nykl et al. (2015) with a more in-depth analysis of the results. The aim of the paper is to identify the results that generalise by using two larger test data sets and two publication databases, one of which is multi-disciplinary. Furthermore, we include other author impact indicators in the evaluation such as a percentile-based indicator R6 (Leydesdorff, Bornmann, Mutz, & Opthof, 2011) and the *PR*-index (Gao et al., 2016), which combines PageRank and a variant of the *h*-index (Hirsch, 2005).

In addition, we analyse the impact that self-citations have on author impact indicators and evaluate different approaches of normalising the author citation graph for PageRank. Lastly, we analyse different approaches of computing impact scores for papers and how these scores should be distributed among co-authors to achieve the best ranking results in ranking well-established researchers.

With this paper, we also present a large test data set consisting of openly available information that can be used to evaluate author impact indicators. The test data comprises author lists of 27 awards handed out to 596 renowned researchers and 1000 authors that received fellowship accreditation by the ACM. We manually matched all researchers in the test data to two publication databases, the ACM's Digital Library (ACM, Inc., 2015) and Microsoft Academic Graph (MAG) (Microsoft, 2017b).

For the evaluation, we focus on variations of the PageRank algorithm (Brin & Page, 1998; Pinski & Narin, 1976) because it is frequently applied to academic citation networks to find important papers (Chen, Xie, Maslov, & Redner, 2007; Dunaiski & Visser, 2012; Hwang, Chae, Kim, & Woo, 2010) and on author citation graphs to rank authors (Dunaiski et al., 2016; Fiala & Tutoky, 2017; Nykl et al., 2015; West, Jensen, Dandrea, Gordon, & Bergstrom, 2013), and has continuously yielded good results as an impact indicator.

We use the average rank as an evaluation measure and use a new methodological approach to estimate the minimum difference required to conclude that rankings are significantly different (Dunaiski, Geldenhuys, & Visser, 2018). Applying this approach, we can compute the significance levels of the differences between two or more rankings. For example, how significant is the difference in the average rank of the authors in the test data when including or excluding self-citations for a certain metric?

With this paper we make the following contributions:

- We make a large test data set available consisting of researchers that won renowned prizes and researchers that are ACM fellows. The author names in these test data sets are matched to author entity identifiers of the ACM and MAG publication databases.
- Based on this test data, we show that using ranking algorithms based on PageRank outperform citation counts as impact indicator of well-established researchers.
- We show that almost all impact indicators are significantly improved by removing self-citations.
- We analyse the effects of different author graph normalisation approaches on the results of PageRank and find that it is more important to normalise the author citation graph than to personalise the PageRank algorithm.
- We find that evenly dividing paper scores among co-authors yields the best results by consistently ranking the authors in our test data higher, independent of which impact indicator is used to compute paper scores.

In this paper, we first review previously published work that uses either awards or fellowship information as test data to evaluate author impact indicators (Section 2). We then provide mathematical definitions of the author ranking algorithms used in this paper, as well as the definitions of the paper credit distribution functions and author citation graph normalisation approaches (see Section 3). In Section 4 we describe the methodology of obtaining and using the test data used for evaluating the indicators, followed by a discussion on the results of our analyses (Section 5).

2. Terminology and related work

PageRank is an algorithm that is computed on graph of vertices which are connected by edges. For example, a paper graph can be used in which the vertices are a set of papers that are connected by citations. In this case, PageRank computes importance scores for papers and we refer to this as a paper-level metric. Alternatively, PageRank can also be computed on an author graph. In this case, PageRank becomes an author-level metric where vertices comprise a set of authors connected by author citations. An author citation occurs between two authors when one author publishes a paper that references a paper by another author.

Fiala et al. (2008) propose a PageRank variation for ranking authors by modifying it to include author collaboration information. They use 15 researchers that won the "ACM E. F. Codd Innovations Award" to evaluate their proposed algorithm on the DBLP database (Schloss Dagstuhl, 2017) and find that the PageRank variation that includes collaboration information tends to perform better than the standard PageRank algorithm.

Nykl et al. (2014) analyse various PageRank approaches and the effects that author graph normalisations and self-citations have on the results of ranking authors. They find that computing PageRank on the paper graph and evenly dividing a paper's score among its co-authors yield the best results. As evaluation data they use two sets of computer science awards, a list of highly cited researchers as produced by Web of Science (WoS), and a list of ACM fellows.

Nykl et al. (2015) test various PageRank personalisations where the PageRank algorithm is initialised with values for each author or paper (Journal Impact Factor on a paper level, *h*-index for authors on the author level, citation counts on the paper level) and explore different paper credit distribution over the order of co-authors of a paper. As publication database they only use journal article entries from the WoS database in the field of computer science between 1996 and 2005 and for fine-grained analyses focus on papers that fall into the "artificial intelligence" and "hardware" categories. As evaluation data they use ACM fellows (1994–2011) and a list of researchers that have won awards from Special Interest Groups for "artificial intelligence" and "hardware design". They find that, using the ACM fellows for evaluation, PageRank on the author graph produces the best results when self-citations are removed and when author personalisation consists of the sum of the journal impact values associated with the authors' papers.

Fiala et al., 2015 use three computer science categories of the WoS database to evaluate 12 different author ranking methods of which nine are PageRank variants. As evaluation data they use a list of editorial board members of the top-10 journals in the fields of computer science based on the Journal Impact Factor. Using this list for evaluations, they find that no PageRank variant outperforms the citation counts of authors. However, when comparing the PageRank variants against 28 "ACM Turing Award" winners and setting PageRank's damping factor to 0.5, they find that PageRank performs slightly better but is still far from outperforming citation counts.

Dunaiski et al. (2016) evaluate various author impact indicators and the author-level Eigenfactor metric (West et al., 2013) using 249 computer science researchers that won innovation and contribution awards on two different databases, the ACM database and the Microsoft Academic Search (MAS) database (the predecessor of MAG). They find that Eigenfactor is the best-suited ranking algorithm to identify high-impact authors.

Gao et al. (2016) combine PageRank and the *h*-index to propose a ranking indicator for authors called *PR*-index. For each paper a PageRank value is computed and used in the *h*-index computation for each author instead of the papers' citation counts. They use papers and cross-citations from the MAS database published between 1992 and 2011 with the keyword "Data Mining". As evaluation data, the authors use researchers that have won the SIGKDD innovation award which consists of only 10 data points. The authors claim that their indicator produces more reasonable results compared to other indicators (citation counts, publication counts, *h*-index, co-author counts, PageRank on author citation graph, and PageRank on co-author graph), since both the popularity and the authority of each publication are considered.

Fiala and Tutoky (2017) use two lists of computer science awards to analyse the performance difference between citation-based and PageRank-based rankings on a computer science subset of the Web of Science publication data. They find that the relative performances depend on which award is used as test data. For instance, they find that citation-based indicators identify recipients of the "ACM E. F. Codd Innovations Award" better, while PageRank-based rankings perform better for the "ACM A. M. Turing Award".

3. Author indicators and algorithms

Using citation counts as an indicator for ranking author importance is the most wide-spread metric and arguably the most intuitive. There are plenty of known caveats of using only citation counts to rank authors. For example, it does not measure productivity in terms of the number of papers published since total citation counts can be skewed by co-authoring a small number of highly cited papers. Furthermore, citation counts of papers can also be inflated by trending popularity in contrast to the intrinsic quality of the work. Martin (1996) distinguishes between research quality, impact and importance, and argues that citation counts best assess a paper's impact.

In this paper our goal is to evaluate how well indicators can rank well-established researchers that have made impacts in their fields. We use cumulative citation counts for authors as a baseline indicator of impact. Let *P* be the set of papers where

$P(a)$ is the set of papers co-authored by author a and where $Cite(p)$ is the number of citations of paper p. The total citation count (Citations) for author a is computed as:

$$\text{Citations}(a) = \sum_{p \in P(a)} Cite(p) \tag{1}$$

The publication count (Papers) of an author a, formally defined as $|P(a)|$, is an indicator that measures the author's lifetime achievements rather than impact and we use this indicator purely for comparative reasons. Similarly, we also give results based on the number of collaborators that authors have (Co-authors).

The h-index, proposed by Hirsch (2005), is defined as:

> An author has an index h if their h most-cited publications have received at least h citations each.

More formally, let $\left\{p_1, p_2, p_3, \ldots | Cite(p_i) \geq Cite(p_{i+1})\right\}$ be an author's set of papers sorted in descending order of citations. The h-index is then computed by stepping through this set and finding the largest value for h such that $h \leq Cite(p_h)$. Like citation counts, the h-index is accumulative in that an author's score does not decrease over time, even if the author does not contribute to the research corpus anymore.

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

> An author has an index 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 $g \leq \frac{1}{g} \cdot \sum_{i \leq g} Cite(p_i)$. Both the h-index and the g-index measures are based on the amount of an author's research output and its impact, and are therefore well-suited for identifying well-established researchers that have made impactful contributions in their fields.

PageRank is an algorithm that computes the importance of a vertex in a graph. It can be computed on any graph or network and follows a Markov chain process to rank the vertices in the graph according to their reachability. The analogy of a random researcher can be used to explain the intuition behind the algorithm, where random researchers are randomly placed on vertices and then follow edges to other vertices. This continues until the random researchers are teleported to new random vertices which is controlled by a teleportation probability $(1-\alpha)$, where α is called the damping factor. If the random researchers reach dangling vertices that have no outgoing edges, they restart their searches on new randomly chosen vertices. The result of PageRank therefore is the likeliness of a random researcher reaching a particular vertex. The better connected a vertex is and the more incoming edges it has from other well-connected vertices, the higher its PageRank score.

PageRank can also be initialised with a *personalisation* vector. This can be used to weight the probabilities of the random researchers starting and restarting at certain vertices. In a paper citation graph we could, for example, skew the probabilities towards papers in a certain field or papers published during a certain time period.

PageRank can also be applied to the author citation graph to compute "importance scores" for authors. In this case the vertices in the graph are authors while edges are author citations. In other words, the weight of the edge from author A to author B is the number of papers co-authored by author A that reference papers co-authored by author B.

More formally, let A be the adjacency matrix of the author citation graph containing n vertices, where A_{ij} is set to the number of times papers of author i have cited papers of author j. For the PageRank algorithm to converge, A must be a left stochastic matrix (each column's sum is 1). Let $\boldsymbol{d}$ be a vector with values $d_i = 1$ if author i is a dangling author (no outgoing citations) in the graph and 0 otherwise. Let $\boldsymbol{r}$ be the personalisation vector which contains the likeliness of a random researcher starting or restarting at a certain author in the graph. PageRank is initialised with $\boldsymbol{x}_0 = \boldsymbol{r}$ and scores for authors are iteratively computed according to Eq. (2) until a predefined precision threshold δ is reached, i.e., $\left\|\boldsymbol{x}_t - \boldsymbol{x}_{t-1}\right\|_1 < \delta$.

$$\boldsymbol{x}_t = \underbrace{\frac{(1-\alpha)}{n} \cdot \boldsymbol{r}}_{\text{Random Restarts}} + \alpha \cdot (A^T + \underbrace{\frac{1}{n} \cdot \boldsymbol{r} \cdot \boldsymbol{d}^T}_{\text{Dangling Vertices}}) \cdot \boldsymbol{x}_{t-1} \tag{2}$$

Implicitly this algorithm defined here adds one edge from each dangling vertex to every other vertex in the graph. The weights associated with these added edges are given by $\boldsymbol{r}$. This is modelled by the "Dangling Vertices" term in Eq. (2), while the first part of the equation, $(1-\alpha)/n \cdot \boldsymbol{r}$, models the distributed placement of random researchers when they restart a search which is controlled by α whose default value is 0.85. In all computations for this paper, we set the precision threshold to $\delta = 10^{-6}$ and limit the number of iterations to a maximum of 100.

The PageRank algorithm can also be computed on journal cross-citation graphs, where the values in the adjacency matrix A_{ij} in Eq. (2) are set to the number of times papers published in journal i have cited papers of journal j. This is the method used by Bergstrom and West (2008) to compute the Eigenfactor scores for journals.

For the Eigenfactor computation, PageRank is personalised with each journal's number of published papers and all journal self-citations are removed. The Eigenfactor score signifies the total influence of a journal. Dividing a journal's Eigenfactor score by its number of papers yields the Article Influence score which is a per-article influence score of a journal. This is similar to the Journal Impact Factor which also computes per-article impact scores for journals. It should be noted that journal cross-citation graphs are usually truncated to only include recent years. The Journal Impact Factor uses the two preceding years to compute journal impact scores for the current year, while the Eigenfactor method uses the preceding 5 years.

Gao et al. (2016) introduced the *PR*-index which uses PageRank to compute a scores for papers after which each paper's score is normalised for the *h*-index computation. Let $PR(p)$ be the PageRank value for a paper p, where p_h is the paper with the highest PageRank value. The rescaled PageRank value $PR'(p)$ is then

$$PR'(p) = PR(p) \cdot \frac{Cite(p_h)}{PR(p_h)} \tag{3}$$

and the final normalised PageRank value $PR_{norm}(p)$ is computed as follows:

$$PR_{norm}(p) = PR'(p) - PR'(p_\ell) \tag{4}$$

where p_ℓ is the paper with the smallest PR' value. Lastly, the *PR*-index for paper p is $\max\{PR_{norm}(p), Cite(p)\}$. These paper scores are then used in the *h*-index computation instead of the papers' citation counts. Gao et al. (2016) argue that by replacing papers' citation counts with PageRank values, the *PR*-index considers both the popularity and impact of an author's publications.

Lastly, we also use the percentile indicator R6 proposed by Leydesdorff et al. (2011). Papers are ranked according to their citation counts separately for every journal and year. Papers are then assigned weights depending on which percentile intervals they belong to. For the R6 indicator, the weights are [1, 2, 3, 4, 5, 6] for the corresponding percentile intervals [0-50, 50-75, 75–90, 90–95, 95–99, 99–100]. For example, a paper receives a score of 5 if it belongs to the top 95% of cited papers for a journal in a specific year. An author's final score is the average of the author's paper scores. We use the approach by Waltman and Schreiber (2013) to compute fair weights for papers across interval boundaries.

3.1. Author self-citations

Self-citation on the author level is a facet that should also be considered when ranking authors. Author self-citations can be an indication of a researcher's or a research group's specialisation in a field where they build upon their own work and that of collaborators (Garfield, 1979; Phelan, 1999). Alternatively, self-citations may be used to manipulate citation rates. Phelan (1999), for example, argues that self-citations should be excluded when performing citation analyses on the author level, but that they do not have a large impact on aggregated levels, such as on the institution or journal levels.

On the basis of this, Aksnes (2003) analysed self-citation rates in the Norwegian scientific literature between the years 1981 and 1996 using a sample of over 46 000 publications. He finds that 21% of all citations are author self-citations and that there exists a strong correlation between the number of authors of a paper and its self-citation rate. Furthermore, he finds that self-citations only contribute to a minor increase in the overall citation counts of multi-authored papers. He also points out that self-citation rates vary significantly between academic disciplines. For example, the self-citation rate in clinical medicine is only 17% while the fields with the highest percentage of author self-citations are chemistry and astrophysics with 31% each. Lastly, Aksnes (2003) concludes that if citation counts are used as research impact indicators, self-citations have a larger influence on the results when the time period of observation after publication is short.

Nykl et al. (2015) and Dunaiski et al. (2016) find that using citation counts with self-citations removed more accurately identifies authors that have won high-impact awards than when the self-citations are included.

As described by Nykl et al. (2014) author self-citations can be removed in two ways. The first approach is to remove a citation between two papers if they have at least one author in common. From the paper citation graph (a) in Fig. 1 the edges *a* and *c* are removed for this approach. It removes a potential citation for each co-author of the cited paper even if some of these co-authors are not on the citing paper. We refer to these citations as *co-author self-citations* and removing all of them is called the "none" method. Alternatively, self-citations can be handled only later when constructing the author graph, by removing all loops (edges from a vertex to itself). We refer to these loops as *author self-citations* and name this method "part" (Fig. 1(c) including the dashed edges). Lastly, the "all" approach includes all self-citations (Fig. 1(b)). We give the self-citation rates of the ACM and MAG databases in Section 5.2 and show the impact that these type of self-citations have on the author ranking algorithms.

3.2. Paper credit distribution over the authors

The question of how to distribute the credit of papers among co-authors has been discussed extensively and many methods of counting multi-authored papers have been proposed. Ideally, one would like to use perfect information about each author's contribution to a paper, but Ajiferuke, Burell, and Tague (1988) show that even interviewing the authors directly may be unreliable. Therefore, the information used as provided on the papers or to the publishers must be used. Trueba and Guerrero (2004) state three principles that should be followed when distributing scores between co-authors. The value of a paper should be shared between authors, divided among authors, and the first author should be credited more than the later authors on the paper. It should be noted that in some disciplines the conventions of author ordering are different. For example, if the ordering of author names is alphabetical, then assigning scores based on author positions would be inappropriate.

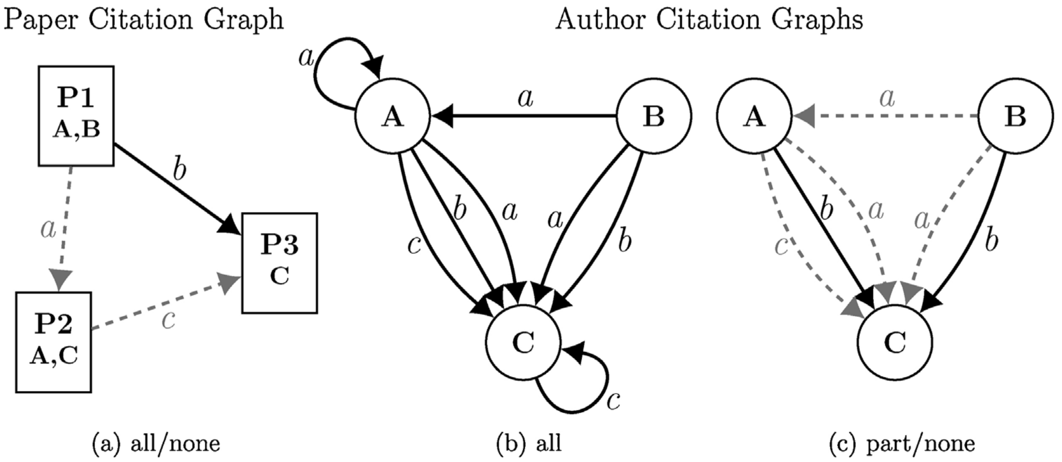


Fig. 1. Illustrative paper citation graph (a) with papers P1, P2 and P3 that are co-authored by the respective authors A, B and C. The graphs in (b) and (c) show how the paper graph is transformed to create author citation graphs where all self-citations are included (b) or where self-citations are excluded (c). The dashed edges in graph (c) indicate co-author self-citations that are omitted when the "none" self-citation approach is used. Alternatively, when only author self-citations are omitted (part), only self-loops are excluded. Similarly for graph (a), the dashed edges are paper self-citations where the citing and the cited paper have at least one author in common.

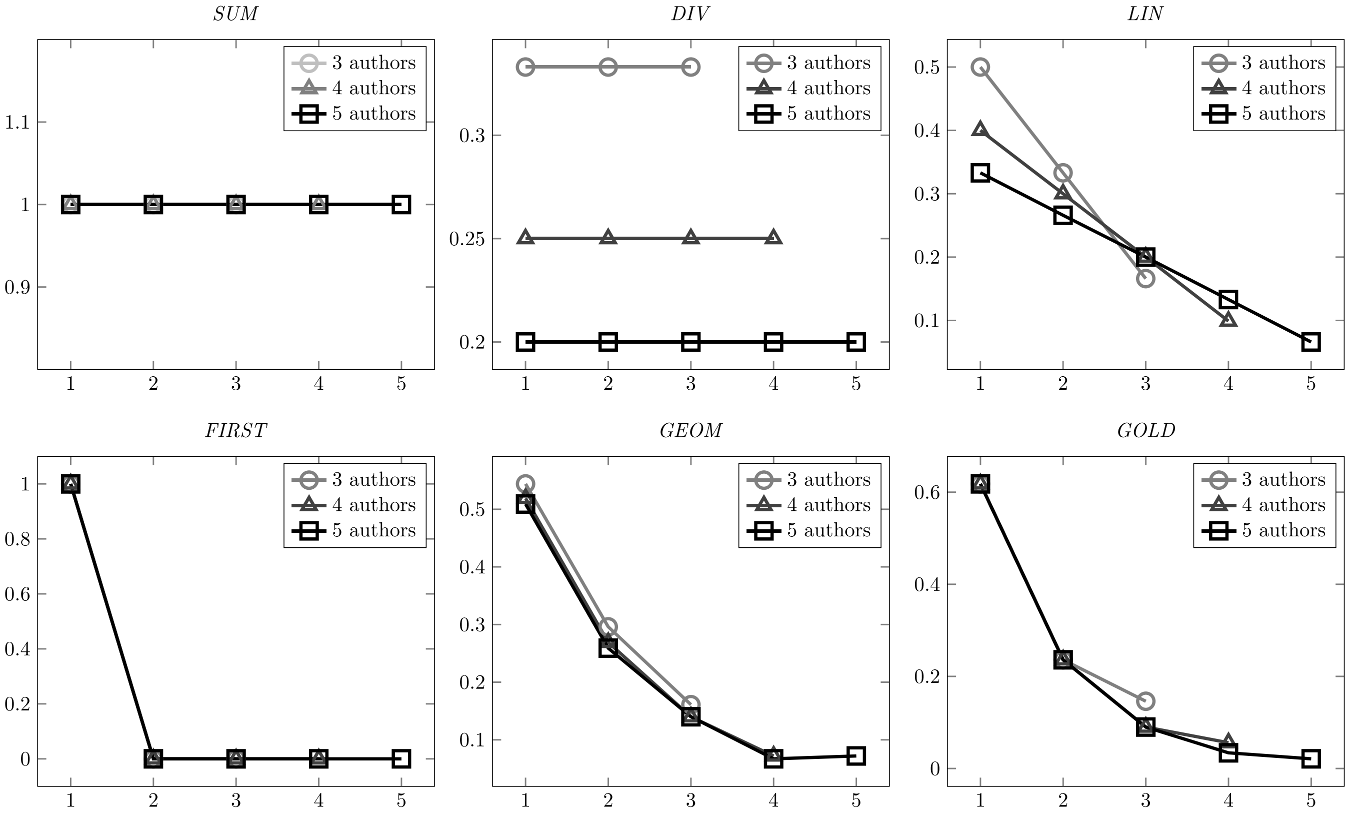


Fig. 2. A figure adopted from Nykl et al. (2015) showing the different paper credit distribution approaches. The graphs show how the paper scores are distributed for papers with three, four and five co-authors. The horizontal axes indicate the author positions in the author lists.

Below, the most common approaches of counting multi-authored papers are reiterated by stating their mathematical definitions. Fig. 2 gives the paper credit distributions of these counting methods by giving examples of papers co-authored by three, four and five authors.

Given a paper p, let $Score(p)$ be the score for the paper as computed by some paper impact indicator, and let $A(p)$ be the set of co-authors. The basic method is to assign the paper score to each author. In other words, the scores of an author's papers are summed to produce the score for an author:

$$SUM(a) = \sum_{p \in P(a)} Score(p) \tag{5}$$

This approach is used by all wide-spread author indicators and only fulfils one of the three principles in that a paper's credit is shared among its co-authors.

Another method is to distribute the score of a paper p evenly between its authors $A(p)$:

$$DIV(a) = \sum_{p \in P(a)} \frac{1}{|A(p)|} \cdot Score(p) \tag{6}$$

This method fulfils two principles since it divides and shares a paper's score among its co-authors.

Let $\text{Pos}(a, p)$ be the position of author a in the author list of paper p. The linear counting method below

$$LIN(a) = \sum_{p \in P(a)} \frac{2}{|A(p)|} \cdot \left(1 - \frac{\text{Pos}(a,p)}{|A(p)|+1}\right) \cdot Score(p) \tag{7}$$

fulfils all three principles.

The first author is often the main driver of a paper and contributes the most work towards its production. Therefore, one could make a supportive argument for a counting method that only accredits the first authors of papers. The "first only" approach is computed as follows:

$$FIRST(a) = \sum_{p \in P(a) | \text{Pos}(a,p)=1} Score(p) \tag{8}$$

This method only fulfils the principle that the first author receives the most credit for a paper.

The geometrical distribution ($GEOM$), proposed by Howard, Cole, and Maxwell (1987), is computed as follows:

$$\begin{aligned} GEOM(a) &= \sum_{p \in P(a)} \left[\lambda^{\text{Pos}(a,p)} \cdot Score(p)\right] \quad \text{where} \\ &\lambda^n + \lambda^{n-1} + \cdots + \lambda^1 = 1 \end{aligned} \tag{9}$$

The gold distribution ($GOLD$), proposed by Assimakis and Adam (2010), computes the paper score attribution according to Formula (10), where $\varphi = 0.618$.

$$\begin{aligned} &GOLD(a) = \sum_{p \in P(a)} [\text{Score}(p) \cdot \Gamma(a)] \\ &\Gamma(a) \begin{cases} 1 & |A(p)| = 1 \\ \varphi^{2 \cdot \text{Pos}(a,p)-1} & \text{Pos}(a,p) = 1, \ldots, (|A(p)| - 1); \quad |A(p)| > 1 \\ \varphi^{2 \cdot \text{Pos}(a,p)-2} & \text{Pos}(a,p) = |A(p)| \end{cases} \end{aligned} \tag{10}$$

The $GOLD$ distribution does not change the ratios of author accreditation except for the last author on a paper. For example, if a paper consists of three or more authors, the first and second authors will always receive 61.8% and 23.6% of the paper's score, respectively. We show the results of applying these different paper credit distribution to the author impact indicators in Section 5.4.

3.3. Normalising the author graph

In the previous section we described how paper scores, after they have been computed using some paper-level impact metric, can be shared among co-authors depending on their positions on papers. Instead of first computing paper scores and then distributing their scores among co-authors, in this section we describe different approaches of how the author citation graph can be normalised before PageRank is used to compute scores for authors.

The author graph on which PageRank is computed is constructed from all the citations between papers. For each citation from paper p_i to paper p_j, with respective authors $a_i \in A(p_i)$ and $a_j \in A(p_j)$, a weight is added to the weights of the corresponding edges in the author graph. These citations can be weighted differently depending on the author lists and the number of references of a paper (which is represented as $\text{od}(p)$ for the out-degree of a paper p). In this paper we use the four different citation weighting approaches described below when constructing author citation graphs. The figures on the right hand side are the results of applying the graph normalisation approaches to the author graph depicted in Fig. 3(b).

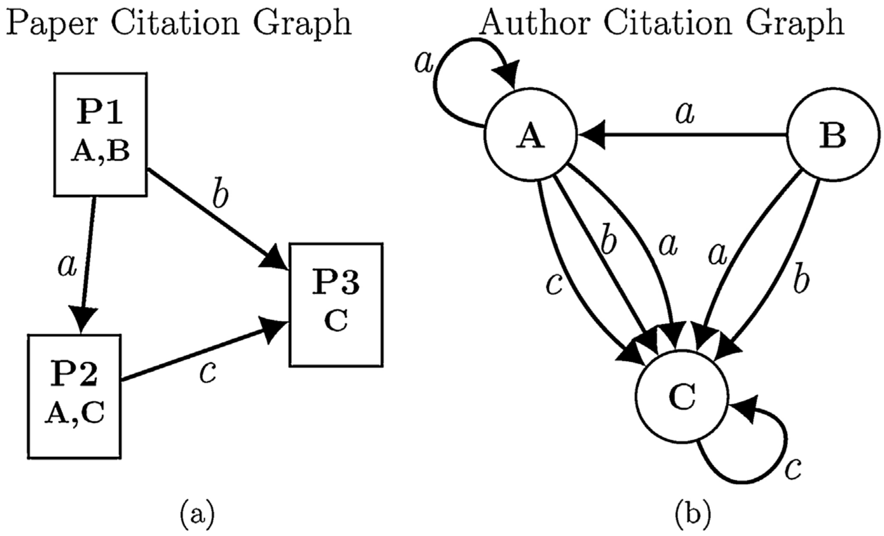
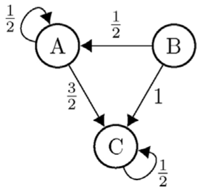
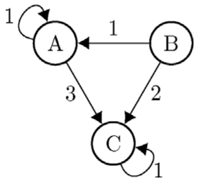


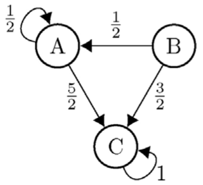
Fig. 3. Illustrative paper citation graph (a) with papers P1, P2 and P3 that are co-authored by the respective authors A, B and C. Graph (b) shows the corresponding author citation graph. A weight of 1 is assigned to the edge weights a, b and c if all citations are treated equally.



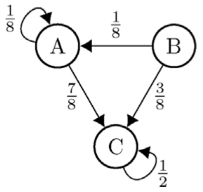
N: $w(a_i, a_j) = \frac{1}{|A(p_i)|} \Rightarrow a = \frac{1}{2}, \quad b = \frac{1}{2}, \quad c = \frac{1}{2}$ This method creates the standard author citation graph, where each citation adds 1 to the in-degree of a cited author. Therefore, the sum of an author's incoming edge weights is equal to the number of papers that cite the author. The in-degrees of the authors A, B and C are therefore 1, 0 and 3.



One: $w(a_i, a_j) = 1 \Rightarrow a = 1, \quad b = 1, \quad c = 1$ Using this method, each added citation contributes 1 to an edge in the author citation graph. In other words, for each citation an author receives, the sum of incoming edge weights is increased by the number of authors on the citing paper. For example, author C is cited twice by paper P1 and once by paper P2, both of which are authored by 2 authors. The in-degree of author C is therefore 6. Similarly, the in-degree of authors A and B are 2 and 0.



OneDivN: $w(a_i, a_j) = \frac{1}{|A(p_j)|} \Rightarrow a = \frac{1}{2}, \quad b = 1, \quad c = 1$ Using this method the citation value of a paper is evenly distributed over the authors that are cited. For example, if the cited paper has two authors, each author gets a half of the citing paper's score attributed.



Eigenfactor: $w(a_i, a_j) = \frac{1}{|A(p_i)| \cdot |A(p_j)| \cdot \mathrm{od}(p_i)} \Rightarrow a = \frac{1}{8}, \quad b = \frac{1}{4}, \quad c = \frac{1}{2}$ This method normalises the citation accreditation over the number of citing authors, the number of cited authors, and the number of references in the reference list of the citing paper. The intuition behind this approach is that a citation from a shorter reference list should count more since it reflects more importance. This method is used by the author-level Eigenfactor method (West et al., 2013).

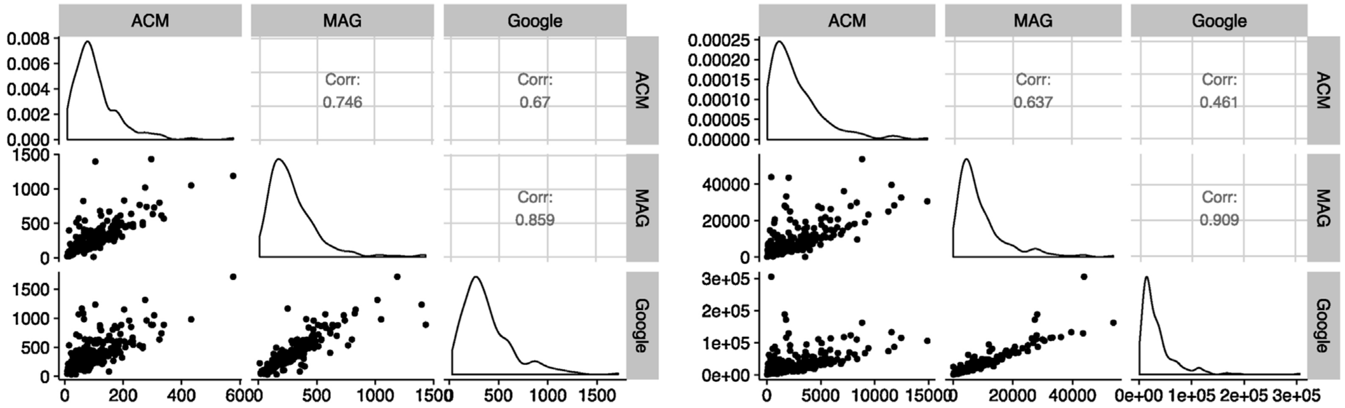
We show the results of using these different author graph normalisation approaches with PageRank in Section 5.3.

4. Methodology

4.1. Data sets used

We use two publication databases for the analyses described in this paper. The first is a copy of the ACM's Digital Library which contains papers up to March 2015 that are published in periodicals and proceedings from the field of computer science. It contains 1 850 715 papers, 1 541 808 authors, and 7 980 995 paper cross-citations. The second database is a snapshot from February 2016 of the Microsoft Academic Graph (MAG) database (Microsoft, 2017a). It is multi-disciplinary and includes 126 909 021 papers, 40 646 689 authors, and 528 682 289 paper cross-citations.

For evaluation purposes of the author impact indicators and ranking algorithms we manually collected two test data sets from various online resources. The first consists of 1000 researchers that are ACM fellows. The ACM fellowship is the recognition of an individual's lasting impact on a field in computer science in terms of technical and leadership contributions,



(a) Correlations between publication counts.

(b) Correlations between citation counts.

Fig. 4. Pearson correlations of the LCA recipients based on the ACM, MAG and Google Scholar databases. Correlation coefficients are shown in the upper triangles, correlation plots are shown in the bottom triangles, while the estimated densities are given on the diagonals.

has influenced the direction of a field, and has to be evidenced by publications, awards, or other publicly recognised artefacts of merit (ACM, Inc., 2017a). Of the 1000 ACM fellows, we use 930 as test data and exclude 70 fellows that have received fellowships for critical inventions or non-academic contributions to their fields such as administrative tasks, work in the industry, or impacts on education or policy. To become a fellowship candidate, researchers first have to be nominated by one of their peers who also submits endorsers. The nominator and the endorsers, collectively, have to be senior enough to make a credible case as to why a candidate's impact merits an ACM fellowship.

The second data set is a list of 596 researchers that have won achievement or lifetime contribution awards (LCA). We considered awards that are handed out by conferences, learned societies, or special interest groups of academic disciplines. Generally, the nomination processes consist of peer nominations and final decisions are taken by dedicated award committees. Of the LCA recipients, only 394 unique authors were matched to the ACM database since 78 researchers received two or more awards and were only counted once. Furthermore, many researchers from fields other than computer science did not have associated entities in the ACM database. For the MAG database, 513 unique authors were matched to database entries.

We manually matched the names of the authors in the test data to the corresponding entries in the ACM and MAG databases. For the ACM database the matching was straightforward since its author-name disambiguation is relatively precise and merging of author entities was not required. Therefore, a single author identifier could be associated with each author in the test data, given that the author has published in the field of computer science and is indexed by the ACM.

The matching for the MAG database was less trivial since its author-name disambiguation and author entity merging is not precise. To find all author entities belonging to a particular author in the test data, we computationally extracted all potential matches. An author entity was considered a potential match if the surname and either the first name or the first initial (for entities without first names) matched an author in the test data. Some author names were nicknames or had varying spellings that do not match the names displayed on papers. Some variation in the order of the first names was observed and first names were often shortened such as 'Tom' for 'Thomas' or 'Bob' for 'Robert'.

For all potential matches we extracted their publication counts, citation counts, affiliations, fields of study, and their papers' keywords from the MAG database. From the list of potential matches, we manually selected entities as matches if a candidate's affiliations matched mostly to the author's affiliations as listed by the ACM and the candidate's fields of study and paper keywords fall predominantly in the same field as the author's field of study for which they received the award. If the candidate has no affiliation information but the field of study or paper keywords fit the author's expertise, we also considered it a match. This mostly occurred for author entities that only had one or two associated papers.

For the 513 unique researchers in the LCA data set that had at least one candidate matched, we found 3.37 database matches on average with a maximum number of 32 matches for a single author. It should be noted that this is a rather conservative number since the matching method we used focused on precision rather than recall.

We also associated the authors in the test data sets with their Google Scholar profiles if they existed and extracted their publication and citation counts in order to obtain an approximation of the correctness of our author entity matching on the MAG database. Fig. 4(a) shows the correlations between the publication counts of the LCA recipients based on the ACM, MAG and Google Scholar databases, while Fig. 4(b) shows the correlations between citation counts.

Note that the MAG and Google Scholar publication and citation counts are highly correlated with Pearson correlation coefficients of 0.86 and 0.91, respectively. If the Google Scholar values are assumed to be accurate, then this indicates that the matching performed on the MAG database is relatively accurate and that the citation counts and paper counts obtained from the MAG database are representative.

The low correlation between the citation counts based on the ACM and the other two databases can be explained by the fact that the ACM database only contains internal citations so that all citation from papers that are not indexed by the

ACM are excluded. The same argument can be used for the relatively low correlation between the paper counts since papers published at venues that are not indexed by the ACM are not included.

4.2. Evaluation

As mentioned in the previous section the entities in the ACM fellows test data set are unique. This is not true for the LCA authors since some authors have won more than one award. Moreover, these two data sets are not disjoint and about 20% of ACM fellows have also won at least one LCA award. Since this overlap is significant, we split the test data in the following way to construct three different test data sets:

ACM fellows is the set of authors that are ACM fellows and have not won a LCA award. This results in a set of 732 entities on the ACM database and 741 on the MAG database.

LCA recipients is the set of authors that have won one or more LCA awards but that do not hold an ACM fellowship. On the ACM (MAG) database this set comprises 195 (318) unique author entities.

INT authors are ACM fellows that have won at least one LCA award. This intersection consists of 198 authors on the ACM database and 189 on the MAG database.

To compare the various impact indicators and algorithms their output scores have to be converted to ranks. The conversion from scores to ranks is done by sorting the scores in descending order. In the case of ties, which happens often with discrete value metrics such as the *h*-index, fractional ranks are assigned which is the average rank between the tied authors. For example, authors with the scores $\{25, 24, 24, 20\}$ would have the corresponding ranks of $\{1, 2.5, 2.5, 4\}$.

It should also be noted that ACM fellows and LCA authors do not necessarily have many published papers or obtained many citations. This is important since we are also interested in how well the ranking algorithms identify entities with relatively few citations.

To evaluate an algorithm's performance of identifying the entities in a test data set, their rank distribution has to be converted into a single-valued performance score. To compute performance scores for algorithms we use the average rank of the authors in the test data sets. We use the average rank as evaluation measure for two reasons. Firstly, the average rank measure has been shown to have high discriminative power to identify significant differences between sparse rank distributions (Dunaiski et al., 2018). Secondly, the average rank is easier to interpret compared to other evaluation measures. For example, given a performance difference X in the average ranks between two ranking algorithms, X indicates that the one algorithm ranked the entities in the test data set on average X ranks higher than the other algorithm.

Voorhees and Buckley (2002) proposed a method to estimate the minimum performance difference required by an evaluation measure to consider two rankings significantly different (at a chosen significance level α). We use this method on an adapted framework for rankings of academic entities to compute the minimum difference in the average rank required to consider two rankings significantly different (Dunaiski et al., 2018). It should be noted that this method is not directly related to statistical significance tests. The results should be interpreted as follows: given a set of rank distributions and a significance level, say $\alpha = 0.05$, a minimum difference X_{min} is estimated. If the average ranks between two rankings differ by more than X_{min}, we can conclude with 95% "confidence" that the algorithm which produced the smaller average rank value is better in identifying the entities in the corresponding test data set. To compute performance scores and compare the results of the different algorithms, only authors can be considered that received a score from each metric. Otherwise, rankings with different population sizes would introduce some bias when comparing the rank distributions of the relevant entities.

Lastly, it should be noted that the author disambiguation on the MAG database is poor. We found multiple author entities associated with the authors in our test data. However, we did not perform author disambiguation and merging on the rest of the author entities in MAG database that are not in our test data. Therefore, to obtain fair results, only a single entity can be used to compute average ranks because we do not know the duplicate entities of other author entities. For each author in the test data, we choose the MAG author entity with the most papers. As a result, some author entities that have papers with high citation counts might be missed. Therefore, we expect the metrics based predominantly on paper counts to perform slightly better on the MAG database compared to the ACM database.

5. Results

In Section 5.1 we give the results of evaluating the author impact indicators using their default configurations, where all citations are included and treated equally. For the PageRank algorithm no personalisation or author graph normalisation is used. After these baselines are established, we analyse the impact that author and co-author self-citations have on the results in Section 5.2.

In Section 5.3 we show how the different author graph normalisations impact the PageRank results of ranking the authors in the test data sets. For example, should the edges between two author nodes be normalised by the number of co-authors on the citing or cited papers?

In Section 5.4 we give the results of the different approaches of distributing paper scores over co-authors. Here, we first compute scores for papers using paper-level impact indicators such as citation counts and more complex ranking algorithms such as computing PageRank on the paper graph. Once the scores for papers are computed, we use the distribution approaches

Table 1
The average ranks of the ACM fellows, LCA recipients, and INT authors on the ACM and MAG databases as produced by the various author indicators with their default configurations. The bottom part of the table shows the corresponding significance levels. Each table cell shows the confidence levels to which the ranking of the row is significantly better from the ranking of the column. The values in the left and right columns correspond to the ACM and MAG database. The rows correspond to the three test data sets: (1) ACM fellows, (2) LCA recipients, and (3) INT authors.

		Co-authors	Papers	R6	h-index	Citations	g-index	PR-index	PageRank
ACM	ACM	21 035	10 902	11 428	8 958	8 422	8 411	7 691	6 042
	LCA	58 192	29 618	29 800	24 352	19 990	24 426	22 745	14 335
	INT	13 144	5 965	4 587	2 495	1 782	2 455	1 809	965
MAG	ACM	1 670 568	460 161	349 009	443 479	401 344	392 350	253 500	–
	LCA	3 601 701	1 269 379	1 488 294	1 542 459	1 571 662	1 414 593	1 256 719	–
	INT	972 797	218 973	124 982	125 723	82 152	102 737	56 753	–

	Co-authors		Papers		R6		h-index		Citations		g-index		PR-index		PageRank	
Co-authors	–															
Papers	*** ** ***	*** ** ***	–			 .		 *		 *		 ,				
R6	*** ** ***	*** ** ***	 *	* ***	–			*		.		,				
h-index	*** ** ***	*** ** ***	* * ***	 ***	*** * ***		–									
Citations	*** ** ***	*** ** ***	*** *** ***	. ***	*** ** ***	 *	, * .	, *	–		 * .	 ,	 .			
g-index	*** ** ***	*** ** ***	*** * ***	. ***	*** * ***	 ,	,	. , ,		 .	–					
PR-index	*** ** ***	*** ** ***	*** * ***	*** ***	*** * ***	* . **	* .	*** * **	,	** * .	, .	** . *	–			
PageRank	*** ** ***	–	*** ** ***	–	*** ** ***	–	*** ** *	–	** * *	–	** ** *	–	*	–	–	

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

discussed in Section 3.2 to obtain ranking scores for authors by, for example, dividing a paper's score evenly between its co-authors.

In Section 5.5 we give the optimal results achieved by each indicator and ranking algorithm after optimising their parameters. Lastly, in Section 5.6 we put our results in the context of previous findings and discuss discrepancies. We also summarise the results that seem to be generally applicable, independent of the test data sets, publication databases, and algorithm parameter settings used.

5.1. Indicators and their default configurations

The top section of Table 1 shows the results of the various indicators with their default configurations where all self-citation are included and no personalisation is used for the PageRank algorithm and the *PR*-index. The average rank is used to measure the algorithms' performance for each of the three test data sets on both the ACM and MAG databases. Therefore, the smaller the average rank value, the higher an algorithm ranks the entities in the test data sets on average.

The bottom section of Table 1 shows the significance levels of the differences between two corresponding rankings. Each table cell shows the levels to which the ranking produced by the metric of the row is significantly different to the ranking of the metric in the column. The left and right values correspond to the ACM and MAG databases. Furthermore, the rows correspond to the three test data sets: (1) ACM fellows, (2) LCA recipients, and (3) INT authors.

For this table and the remainder of this section, we only show significance values in the cases where $\alpha < 0.15$ and the average rank value produced by the metric of the row is smaller than the one produced by the metric of the column. Therefore, a significance value indicates the confidence that the metric of the corresponding row produced a significantly better result compared to the metric of the column.

For example, the first row ('Co-authors') does not contain any significance values which means that the Co-authors metrics does not produce significantly better results compared to any other metric. However, the cell in the second row ('Papers') and the first column ('Co-authors') contains six significance values. The first value in the left column ('***') corresponds to the ACM fellows test data set and the ACM database. It indicates that the difference in the average rank produced by paper counts and co-author counts is significantly different with 99.5% confidence. In other words, there is a 0.5% probability that the significant difference in the average rank produced by these two ranking algorithms is observed by chance.

As another example, consider the cell of the row 'Citations' and the column '*PR*-index' which only contains a single significance value ('.'). This means that the average rank produced by citation counts is significantly better than the one produced by *PR*-index only for the LCA recipients on the ACM database and with 90% confidence.

Considering the average ranks listed in Table 1, PageRank performs the best in identifying the authors in all three test data sets on the ACM database. However, only for the ACM fellows is the difference in the average rank a significant improvement over the *PR*-index. After PageRank, the *PR*-index performs the best. On the ACM database however, citation counts outperform *PR*-index when ranking the LCA recipients and the INT authors. However, only for the LCA recipients is the difference significant ($\alpha < 0.1$). Gao et al. (2016) state that the *PR*-index achieves better results in ranking researchers than citation

Table 2
The average number of citations of the entities in the test data sets, as well as their average author and co-author self-citation rates. For comparison, the row 'All authors' shows the corresponding values when all entities of the entire databases are considered.

	ACM database			MAG database		
	Citations	Author self-citations	Co-author self-citations	Citations	Author self-citations	Co-author self-citations
ACM fellows	1569.83	6.97%	11.74%	4155.17	4.74%	8.22%
LCA recipients	1208.00	5.31%	9.07%	4382.39	3.59%	5.95%
INT authors	3080.52	5.82%	9.84%	9456.84	4.30%	6.93%
All authors	30.65	8.50%	22.27%	66.53	1.99%	9.40%

counts and PageRank on the author citation graph. We find that the *PR*-index, in most cases, does identify high-impact researchers better than citation counts. However, we do not find significant differences between the *PR*-index and PageRank, except for the ACM fellows on the ACM database, where PageRank performs better ($\alpha < 0.005$).

The *g*-index and citation counts perform relatively similar. However, citation counts does perform significantly better when the INT authors are used on both databases and the LCA recipients on the ACM database. Both these metrics perform significantly better than the *h*-index in most cases.

Citation counts perform better than publication counts, co-author counts, and the *h*-index, except when publication counts are used to rank the LCA recipients on the MAG database. Both publication and co-author counts do not measure an author's impact directly and rather reflect their life-time achievement and the extent of their contributions. Therefore, based on these results, citation counts only significantly outperform one author impact indicator, the *h*-index.

The results for PageRank on the MAG database are missing due to the impracticably high computation costs when PageRank is computed on the author citation graph. Compared to the paper citation graph, the memory requirements for computing PageRank on the author graph is a magnitude higher (over 1.5 terabytes compared to around 150 gigabytes) since it is a very densely connected graph. Therefore we chose to compute PageRank only once on the MAG database. For the one-time computation we elected to use the default author-level Eigenfactor parameters. Since these parameters are not the default configuration of PageRank, the results are not listed in this section and instead are discussed in Section 5.5.

5.2. *The impact of self-citations*

In this section we investigate how self-citations affect the results of the various indicators. Table 2 lists the average number of citations, the author self-citations rates, as well as the co-author self-citation rates of the entities in the test data sets. The last row shows the corresponding values for the entire publication databases.

Considering the self-citation rates of the ACM fellows on the ACM database, their average author self-citation rate is 6.97% which increases to 11.74% when all co-author self-citations are included. The LCA recipients and INT authors have similar self-citation rates. The overall average self-citation rate for all authors in the ACM database is 8.50% and 22.27% for co-author self-citations. Considering the MAG database, the overall author and co-author self-citation rates are 1.99% and 9.40%, which are considerably smaller compared to the ACM database.

Author entity merging was only performed for the authors in the test data sets. Therefore, the self-citation rates based on the entire MAG database have to be interpreted conservatively. The true self-citation rates are expected to be slightly higher. However, the average co-author self-citation rate using the MAG database (9.40%) is still higher than that of the authors in the test data for which we did merge author entities. Therefore, we can conclude that the authors in the test data have lower self-citation rates on average than the authors in the databases.

Table 3 shows the average ranks achieved by the indicators on the ACM database with different self-citation strategies. In parentheses we give the delta in the average rank compared to the baselines where all self-citations are included. For these PageRank computations no personalisation was used and all citations were treated the same (i.e., the **One** author graph normalisation approach). Again, the results for the PageRank algorithm are not available for the MAG database, since the default author-level Eigenfactor metric uses authors' publication counts for personalisation. From Table 3 we can see that the results of all metrics improve by removing self-citations. The improvement is larger when all co-author self-citations are removed ('none') compared to only removing author self-citations ('part'). In fact, the difference in the average ranks between including all self-citations and only omitting author self-citations is not significant in many cases. However, the difference between "all" and "none" is significant in most cases. The exceptions are the *PR*-index when the ACM fellows are used and the PageRank algorithm when the LCA recipients are used.

Table 4 is similar to the previous table and shows the results when the MAG database is used. The results are very similar and all methods improve when omitting all co-author self-citations. However, for the ACM fellows the average rank increases for the *h*-index, the *g*-index, and citation counts when only omitting author self-citations. The same is true when citation counts are used to rank the INT authors. However, in each case the difference in the average rank is not significant.

When only considering the citation counts metric, we find that removing all co-author self-citations performs the best ($\alpha < 0.005$), independent of the test data and publication database used. The improvement by removing all co-author self-citations compared to only removing author self-citations is also significant ($\alpha < 0.005$), except for the INT authors on the ACM database. Nykl et al. (2014) use two small sets of authors that won contribution awards and find conflicting results

Table 3
The average ranks of the ACM fellows, LCA recipients, and INT authors on the ACM database produced by the various ranking algorithms with different self-citations strategies. The values in parentheses show the difference to the baseline of the same method where all self-citations are included ("all"). The rows "part" indicate the omission of author self-citations, while "none" indicate the omission of all co-author self-citations. The significance levels of the differences between two average ranks are indicated by the brackets.

Indicator	Self-Cites	ACM				LCA				INT			
Co-authors	n/a	21,035				58,192				13,144			
Papers	n/a	10,902				29,618				5,965			
R6	all	11,428		***		29,800		.		4,587		***	
	none	11,056	(-371)			29,440	(-361)			4,459	(-127)		
h-index	all	8,958		,		24,352		***		2,495		,	
	part	8,725	(-233)		***	23,281	(-1,071)		***	2,331	(-164)		***
	none	8,396	(-562)	**		22,060	(-2,293)	***		2,211	(-284)		
Citations	all	8,422				19,990		**		1,782			
	part	8,278	(-144)		***	19,279	(-711)		***	1,726	(-56)		***
	none	8,018	(-404)	***		18,332	(-1,658)	***		1,699	(-83)		
g-index	all	8,411				24,426		***		2,455			
	part	8,328	(-83)		***	23,215	(-1,212)		***	2,330	(-125)		***
	none	8,066	(-345)	*		22,271	(-2,155)	***		2,236	(-218)		
PR-index	all	7,691				22,745		***		1,809		***	
	none	7,651	(-40)			19,275	(-3,470)			1,567	(-243)		
PageRank	all	6,042				14,335				965			
	part	6,019	(-23)		***	14,252	(-83)			949	(-16)		**
	none	5,710	(-332)	***		13,986	(-349)			912	(-52)		

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

Table 4
The average ranks of the ACM fellows, LCA recipients, and INT authors on the MAG database produced by the various ranking algorithms with different self-citations strategies. "part" indicates the omission of author self-citations and "none" indicates the omission of all co-author self-citations. The values in parentheses show the difference to the same method where all self-citations are included. The significance levels of the differences between two average ranks are indicated by the brackets.

Indicator	Self-Cites	ACM				LCA				INT			
Co-authors	n/a	1,670,568				3,601,701				972,797			
Papers	n/a	460,161				1,269,379				218,973			
R6	all	349,009		***		1,488,294		.		124,982		***	
	none	341,943	-7,066			1,470,278	-18,016			122,287	-2,695		
h-index	all	443,479				1,542,459		*		125,723			
	part	443,643	163		***	1,526,333	-16,126		***	122,685	-3,038		,
	none	425,256	-18,224	***		1,488,018	-54,441	***		120,541	-5,182		
Citations	all	401,344				1,571,662				82,152			
	part	404,351	3,007		***	1,568,497	-3,165		***	82,528	376		***
	none	389,856	-11,488	***		1,522,627	-49,035	***		78,944	-3,208	***	
g-index	all	392,350				1,414,593				102,737			
	part	392,864	515		***	1,405,441	-9,152		***	101,650	-1,087		***
	none	380,206	-12,144	***		1,373,025	-41,568	***		96,454	-6,283	***	
PR-index	all	253,500		***		1,256,719		***		56,753			
	none	238,227	-15,273			1,196,892	-59,828			53,742	-3,011		

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

when using citation counts as an indicator. For the one award they find that including all citations performs the best, while for the other, only removing author self-citations performs the best. However, when using a larger test data set consisting of ACM fellows, they find that removing all co-author self-citations performs the best which we confirm using both the ACM and the MAG databases.

5.3. Results of PageRank with different graph normalisations

In this section we focus on the PageRank algorithm when applied to the author citation graph and how varying its parameters can influence the ranking results. Table 5 shows the results of comparing the different author graph normalisation approaches. For a single paper-level citation, the **One** approach adds an edge (with weight 1) for every citing author. The **N** approach normalises these edges by the number of citing authors, while the **OneDivN** approach normalises the edges by the number of cited authors. Lastly, the **Eigenfactor** approach normalises the edges by both the number of citing and cited authors, as well as the number of references of the citing paper.

The significance values in Table 5 indicate that the normalisation approach of the row is significantly better than the approach of the column. Each table cell is split into three columns and three rows. The columns correspond to the three test data sets (left to right: ACM fellows, LCA recipients, INT authors). The rows correspond to the three self-citation strategies

Table 5
Comparison of the different author graph normalisations used by the PageRank algorithm. The table shows the corresponding significance levels when comparing normalisation approaches. Each table cell shows the confidence levels to which the ranking of the row is significantly different from the ranking of the column. The columns in a table cell correspond to the three test data sets (left to right: ACM fellows, LCA recipients, INT authors). Similarly, the rows correspond to the three self-citation strategies (top to bottom: all, part, none). For each combination, four significance values are shown which correspond to four personalisation strategies (top to bottom: none, *h*-index, paper counts, citation counts).

	one			N			oneDivN			eigenfactor		
one	–											* * *
												* . .
												* * *
N	* * * *	. . . .	. .									* * *
	* * * *	. . . .	. .									* * *
	* . . *	. . . .	, ,									* * * *
oneDivN	* * * *	* * * *	. . . *	* * * *	* * * *	 , , .	–					* * * *
	* * * *	* * * *	* * * *	* * * *	* * * *	. . . *						* * * *
	* * * *	* * * *	 .	. * * *	, , ,	 ,						* * * *
eigenfactor	* * * *	* * * *		* * * *	* * * *		. * * *	 , .		–		
	* * * *	* * * *	 .	* * * *	* * * *		. * * *	 , .				
	* * * *	* * * *		* * * *	* * * *		. * * *	. . . .				

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

(top to bottom: all, part, none). Lastly, each combination of normalisation approach and self-citation strategy is associated with four significance values which correspond to four different personalisation approaches for PageRank (top to bottom: none, *h*-index, paper counts, citation counts).

From the table we can see that the self-citation strategy and the personalisation approach do not have a large influence on the ranks. The predominant factor is the author citation graph normalisation. Using the **One** approach performs the worst. This is expected since the impact of a paper should not depend on the number co-authors of the citing papers. The **Eigenfactor** normalisation approach performs the best except for the INT authors where it performs the worst. It is unclear why the **Eigenfactor** approach performs poorly for this particular test data set. For completeness, Table A.9 in Appendix shows the results similar to Table 5 but where the different personalisation strategies are directly compared to each other.

5.4. *Different paper score distributions over co-authors*

Instead of computing author impact indicators on the author citation graph directly, one can first compute importance scores for papers and then distribute these scores over the papers' co-authors. In this section, we turn to the question of which distribution approach to use for identifying well-established researchers. As described in Section 3.2, the approaches either distribute the score equally over all authors (*SUM*, *DIV*) or in decreasing order based on the authors' positions (*FIRST*, *LIN*, *GEOM*, *GOLD*).

Using citation counts as a metric for paper scores and the ACM fellows as test data, we show the performance of these approaches in Fig. 5. The three different self-citation variants are also considered, where the results of using all citations are indicated in black bars, while the dark and light grey bars indicate the results of omitting author self-citations (part) and all co-author self-citations (none), respectively.

From Fig. 5 we can see that only giving credit to the first author (*FIRST*) performs the worst, while the best performing approach divides the paper credit evenly across all co-authors (*DIV*). According to these results, not all three principles stated by Trueba and Guerrero (2004) hold true for the optimal paper credit distribution for well-established researchers (i.e., that the first author should be credited more than the later authors). However, the principle that paper scores should be divided among co-authors holds true since *GOLD*, *GEOM*, *LIN*, and *DIV* perform significantly better then *SUM*.

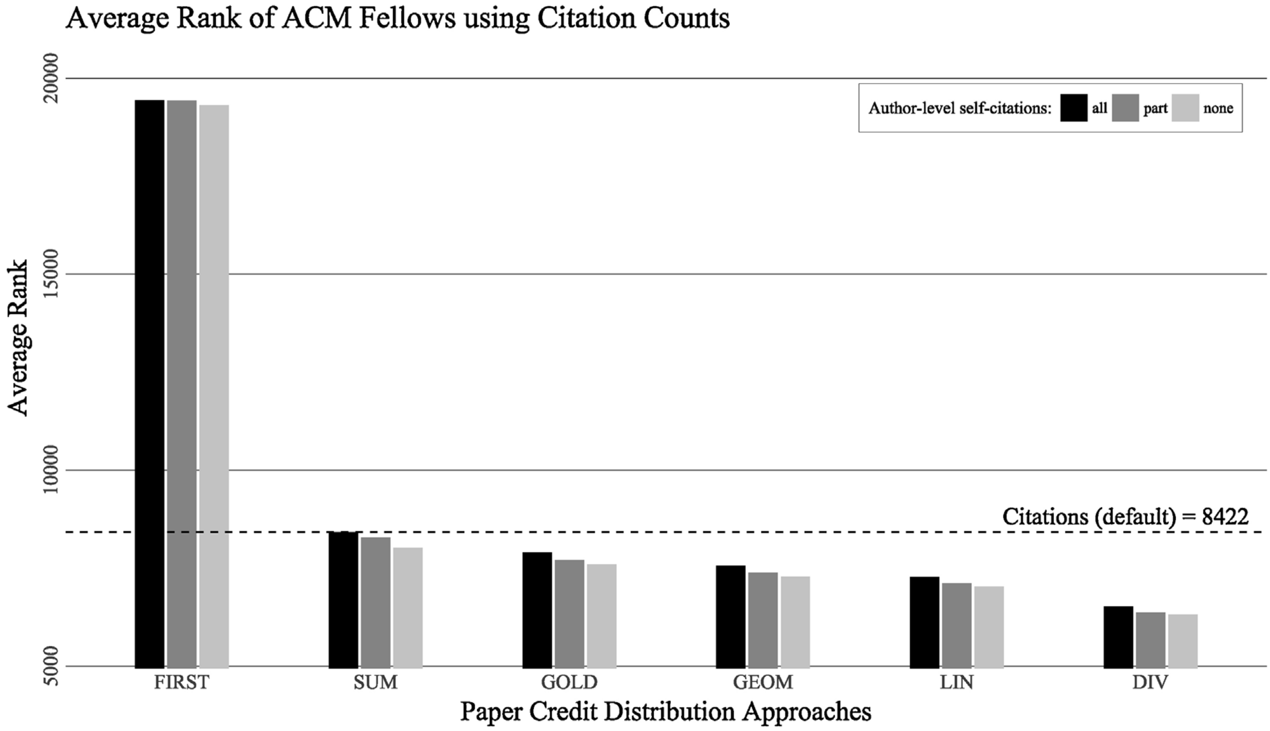


Fig. 5. The average rank of ACM fellows when using citation counts with different paper credit distributions over co-authors computed on the ACM database. The three different self-citation variants are indicated by the differently shaded bars. The reference line "Citations (default)" shows the average rank when using citation counts including all self-citations.

Table 6
Comparison of the different score distribution approaches where papers' citation counts are used as paper impact metric which are distributed among co-authors. The table shows the corresponding significance levels when comparing two approaches. Each table cell shows the confidence levels to which the ranking of the row is significantly different from the ranking of the column. The values in the left and right columns correspond to the ACM and MAG database. The rows correspond to the three test data sets: (1) ACM fellows, (2) LCA recipients, and (3) INT authors.

	FIRST		SUM		GOLD		GEOM		LIN		DIV	
FIRST	–											
SUM	*** ** ***	*** * ***	–									
GOLD	*** ** ***	*** * ***	* * .	*** * ***	–							
GEOM	*** ** ***	*** * ***	* * *	*** * ***	*	.	–					
LIN	*** ** ***	*** * ***	* * *	*** * ***	* .	* .	 , ,	.	–			
DIV	*** ** ***	*** * ***	*** * *	*** * ***	** * ,	*** *	* *	** *	* .	* . .	–	

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

However, this is expected since most ACM fellows are well-established researchers with a long research career and therefore are expected to frequently appear on the last position of author lists. The average time for the researchers to become ACM fellows since their first publication is 23 and 28 years, based on the ACM and MAG databases. Furthermore, we found that 62.92% (ACM) and 64.85% (MAG) of author occurrences of the ACM fellows are on the last position if single-authored papers are excluded.

Table 6 gives the significance levels for differences in the average ranks when comparing the various score distribution approaches (all self-citations included). Each table cell shows the confidence levels to which the ranking with the approach of the row is significantly better than the ranking with the approach of the column. The values in the left and right columns correspond to the ACM and MAG databases and the rows correspond the three test data sets.

The significance values in Table 6 show that the results are similar independent of which test data set and database is used. *FIRST* is the worst performing approach followed by *SUM*, *GOLD*, and *GEOM*. However, the *GOLD* and *GEOM* approaches are very similar. Only for the ACM fellows are they significantly different ($\alpha < 0.05$ for the ACM database and $\alpha < 0.1$ for the MAG database). The *LIN* approach only significantly improves over *GEOM* in three of six cases. The *DIV* approach performs better than *LIN* in almost all cases with a minimum confidence level of 90%. The exception is for the INT authors on the ACM database where no significant difference is found.

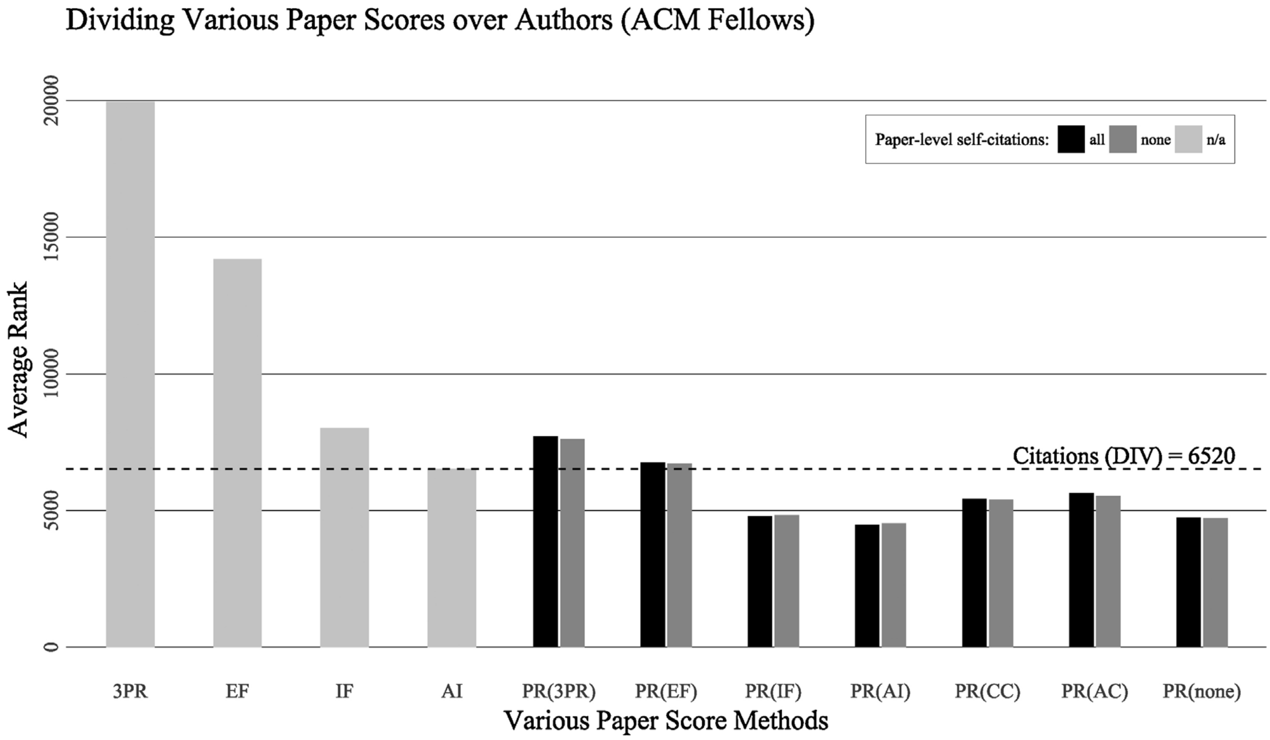


Fig. 6. The average ranks achieved by various paper ranking methods evenly divided between the co-authors using the ACM fellows as test data on the ACM database. The labels "all" and "none" indicate whether methods include or exclude paper-level self-citations. The label "n/a" indicates that paper-level author self-citations do not apply since the paper scores are directly computed using the journal cross-citation graph. The references line "Citations (DIV)" shows the average rank produced by using citation counts as the paper scores and evenly dividing their scores across co-authors.

Following Nykl et al. (2015) we also computed scores for papers using various paper ranking methods and then distributed these paper scores between the corresponding co-authors using the paper credit distributions as described in Section 3.2. We used the following paper ranking methods to obtain scores for papers:

PageRank with paper-level personalisations: *PR*(*CC*), *PR*(*AC*) and *PR*(*none*) indicate the results of the PageRank computations on the paper citation graph with different paper personalisation strategies where *CC*, *AC* and *none* indicate personalisations with a paper's citation count, a paper's author count, and no personalisation, respectively.

PageRank with journal impact personalisations: For the following approaches journal impact scores are associated with each paper which are used for the PageRank personalisation. As journal impact indicators we used the Journal Impact Factor (*IF*) (Garfield, 1972), a 3-year PageRank score for journals (3*PR*), the Eigenfactor (*EF*) and its Article Influence (*AI*) score (Bergstrom, West, & Wiseman, 2008). Therefore, *PR*(*IF*), *PR*(3*PR*), *PR*(*EF*) and *PR*(*AI*) indicate the PageRank computations for papers that are personalised, respectively, with the Impact Factor, 3-year PageRank, Eigenfactor, and Article Influence scores of the associated journals.

Journal impact scores *IF*, 3*PR*, *EF* and *AI* are four approaches in which the paper scores are directly replaced by their journals' Impact Factor, 3-year PageRank, Eigenfactor, and Article Influence scores in which they were published.

It should be noted that the journal impact values are computed for each year. Therefore, a paper's associated journal impact value is the one computed for the year that the paper was published. However, the journal impact values are not normalised over the entire time span of all publication years. The Eigenfactor and Impact Factor for journals are computed using their default time periods of 5 years and 2 years, respectively. Lastly, the PageRank algorithm on the journal citation graph is computed using the default damping factor of 0.85 and including all journal self-citations, while for the Eigenfactor computations journal self-citations are excluded.

Using the ACM database and the ACM fellows as test data, Fig. 6 shows the results of the different paper score algorithms and dividing the score evenly across a paper's co-authors. We also computed the average ranks using the other paper credit distribution approaches but chose to omit the results since the *DIV* method achieves the best results independent of the paper score method used. Since the paper scores are computed on the paper level, author and co-author self-citations are identical. Therefore, in Fig. 6 the methods in which self-citations are omitted are indicated as "none" and methods where all citations are counted are labelled as "all". The methods in which the papers' journal impact scores are used directly, paper self-citations do not exist and are therefore shown as "n/a".

We can see that the methods in which the journal impact values are directly associated with papers are generally worse performing than when the values are used as personalisation scores for the PageRank algorithm on the paper citation

Table 7
The results of evaluating the various author impact indicators after optimising their parameters on the ACM database. The column "Parameters" shows the optimal parameter list (self-citation strategy, damping factor, personalisation strategy, graph normalisation technique) for each indicator. The last three rows are the results of the methods where paper scores are computed on the paper graph and evenly divided (*DIV* approach) among the corresponding co-authors. The column "Sign." shows the significance values associated with the improvements in the average ranks of successive metrics.

Indicator	Parameters	Score	Sign.
Co-authors	–, –, –, –	26 087	***
Papers	–, –, –, –	13 277	
R6	none, –, Citations, –	13 082	***
h-index	none, –, –, –	9 676	.
g-index	none, –, –, –	9 502	, .
Citations	none, –, –, –	8 694	
PR-index	none, 0.85, none, –	8 595	***
PageRank	none, 0.9, none, eigenfactor	5 169	
Papers *DIV*	–, –, –, –	10 320	***
Citations *DIV*	none, –, –, –	6 939	***
PageRank *DIV*	none, 0.95, none, –	5 616	

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

graph. Of the methods that directly use journal impact values, *AI* performs the best with an average rank of 6514 which is significantly better than *IF* with 8012 ($\alpha = 0.005$).[1]

Using PageRank on the paper citation graph with the journals' Article Influence scores as personalisation (*PR*(*AI*)) performs the best overall with an average rank of 4474 and 4525 when self-citations are included and excluded. PageRank with no personalisation (*PR*(*none*)) and *PR*(*IF*) produce very similar average rank values where the differences to *PR*(*AI*) are not significant ($\alpha > 0.15$).

Comparing these results to Nykl et al. (2015), we can confirm that using PageRank on the paper citation graph and personalising it with journal impact values is the best approach and that directly using journal impact values generally yields the worst results. Furthermore, we find that the impact of paper-level self-citations is negligible and that it is more important to personalise PageRank with appropriate personalisation vectors. It should be noted that both 3*PR* and *EF* metrics compute overall impact scores for journals whereas *IF* and *AI* compute per-article journal influence scores. Using these per-article journal influence scores in the paper-level PageRank computations yield the best results.

5.5. Parameter optimisation

To fairly compare the performance of the various indicators with optimised parameters but without overfitting the indicators we merged the test data sets and split the entities into 10 stratified folds. On the ACM and MAG databases the test data sets comprise 1125 and 1248 author entities, respectively. For the ACM fellows, we stratified the authors across the folds such that the years in which the authors received their fellowships are evenly distributed. Similarly for the LCA recipients, we stratified the authors across the folds such that the awarding venues and years are evenly distributed between them.

We use 10-fold cross validation for all algorithms where 9 folds are used for training and the hold-out fold is used for testing. We report the average result of 10 iterations where each fold is used exactly once for testing. Except for the damping factor of the PageRank algorithm which is optimised, all other parameters (self-citation approaches, graph normalisations, personalisations) are treated as hyper-parameters. We optimised the damping factor in the range of 0.05–0.95 with intervals of 0.05.

Table 7 shows the optimal results obtained for each indicator and ranking algorithm when the ACM database is used. Table 8 shows the analogous results for the MAG database. Both tables consist of two parts where the top eight indicators are based on the author graph, while the bottom three indicators use paper scores that are computed on the paper graph and then evenly divided among respective co-authors. The column "Sign." gives the significance values for the metrics where a significant improvement in the average rank is observed. Since the metrics in both parts of the table are sorted in descending order of the achieved score, a corresponding significant improvement by a method also applies to all metrics listed above. We omit redundant significance values for space reasons. For example in Table 7, the difference in the average rank between the *h*-index (9676) and the R6 metric (13 082) is significant with $\alpha < 0.005$. It follows that the *h*-index is also significantly different to 'Papers' and 'Co-authors' to at least the same significance level.

Considering only the results of the ACM database in Table 7 for now, co-author counts (26 087), publication counts (13 277), and the R6 metric (13 082) are the worst performing indicators. Both the *h*-index (9676) and the *g*-index (9502) are a significant improvement ($\alpha < 0.005$) but no significant difference is found between them. Citation counts (8694) perform slightly better than the *h*-index ($\alpha < 0.1$) and the *g*-index ($\alpha < 0.15$).

Gao et al. (2016) did not perform any type of optimisation of their proposed *PR*-index and therefore we evaluated different damping factors, personalisations and self-citation strategies when computing the PageRank values on the paper graph used

[1] All significance values for all test data sets on both the ACM and MAG databases are given in Table A.10 in Appendix.

Table 8
The results achieved by the various author impact indicators after optimising their parameters using the MAG database. The columns "Parameters" and "Sign." are analogous to the columns in Table 7.

Indicator	Parameters	Score	Sign.
Co-authors	–, –, –, –	2 056 963	***
h-index	none, –, –, –	649 909	
Citations	none, –, –, –	631 410	
Papers	–, –, –, –	629 830	,
R6	none, –, Citations, –	596 186	**
g-index	none, –, –, –	590 212	*
PR-index	none, 0.85, *EF*, –	448 807	***
PageRank †	part, 0.85, Papers, eigenfactor	320 628	
Citations *DIV*	none, –, –, –	430 616	.
Papers *DIV*	–, –, –, –	361 386	*
PageRank *DIV*	none, 0.9, *EF*, –	265 302	.

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.
[†]The damping factor was not optimised on the MAG database due to the high computational costs associated with the dense author cross-citation graph.

for the *PR*-index calculations. The personalisation strategies we used for optimising *PR*-index are the ones also used with PageRank in Section 5.4 (i.e., *CC*, *AC*, *none*, *IF*, 3*PR*, *EF*, *AI*). *PR*-index's optimal result (8595) is obtained when α is set to 0.85, no personalisation is used, and all co-author self-citations are excluded. However, it is not a significant improvement over citation counts.

When using PageRank directly on the author citation graph, the best result (5169) is obtained when setting the damping factor to 0.9, using no personalisation, excluding all self-citations, and applying the **Eigenfactor** graph normalisation approach.

For the three metrics that are based on paper impact scores, we again find that using the *DIV* approach yields the best results. We list three different paper impact methods. As a baseline we show the results of Papers *DIV* where all papers are assigned a score of 1. In addition, we show the results of using citation counts for papers and PageRank on the paper graph.

Papers *DIV* performs the worst (10 320), followed by Citations *DIV* with an average rank of 6939. Using PageRank on the paper citation graph and evenly dividing paper scores (PageRank *DIV*), the best result (5616) is obtained when removing all self-citations, no personalisation is used, and the damping factor is set to 0.95. It should be noted that each metric performs the best when all self-citations are removed.

Table 8 shows the results for the MAG database. In general, the results are similar, however, a few differences are observed. Firstly, the R6 metric, paper counts, and the *g*-index perform better than citation counts. Secondly, the *PR*-index and PageRank *DIV* produce the best results when PageRank is personalised with the Eigenfactor journal impact values (PR(*EF*)) on the paper level.

Table 8 also shows the results of the one-time computation of PageRank on the author citation graph using the MAG database. It uses the **Eigenfactor** graph normalisation approach, author self-citations are excluded (part), and the damping factor is set to 0.85. With this parameter combination PageRank achieves an average rank of 320 628. Even though its damping factor was not optimised, PageRank perform significantly better ($\alpha < 0.005$) compared to the other eight indicators computed on the author graph.

PageRank performs better than citation counts regardless of whether it is directly computed on the author citation graph or first on the paper citation graph where paper scores are then evenly divided among co-authors (PageRank *DIV*). On the author citation graph, the improvement is more significant ($\alpha < 0.005$) than on the paper citation graph ($\alpha < 0.05$).

On the ACM database, PageRank on the author graph performs the best but the difference to PageRank *DIV* is not significant ($\alpha > 0.15$). On the MAG database, PageRank *DIV* achieves the best result and is significantly better than PageRank on the author graph ($\alpha < 0.15$). However, it should be noted that in this case the damping factor of the author-level PageRank algorithm was not optimised. In general, removing all co-author self-citations improves the performance of all indicators.

5.6. Discussion

In the following discussion, α refers to the damping factor of the PageRank algorithm and not to the significance level. Fiala and Tutoky (2017) found that citation-based metrics identify recipients of "ACM E. F. Codd Innovations Award" better, while PageRank-based rankings perform better for the "ACM A. M. Turing Award". In this paper, we did not distinguish between different types of awards and therefore we cannot corroborate these findings. However, our results showed that if a larger number of different awards are merged and used as test data on different databases, PageRank-based metrics always performed better.

Nykl et al. (2014) found that the overall best PageRank approach for ranking award winners is to compute PageRank on the paper citation graph personalised with author counts, removing all self-citations, and to evenly distribute paper scores among co-authors. We cannot confirm the exact combination of parameters but our results do indicate that paper-level self-citations should indeed be removed and that paper scores should be evenly shared among co-authors.

Nykl et al. (2015) also found that the overall best approach to ranking high-impact authors is to use a paper's journal impact score for personalisation of the paper-level PageRank computations. We can confirm this observation. In general,

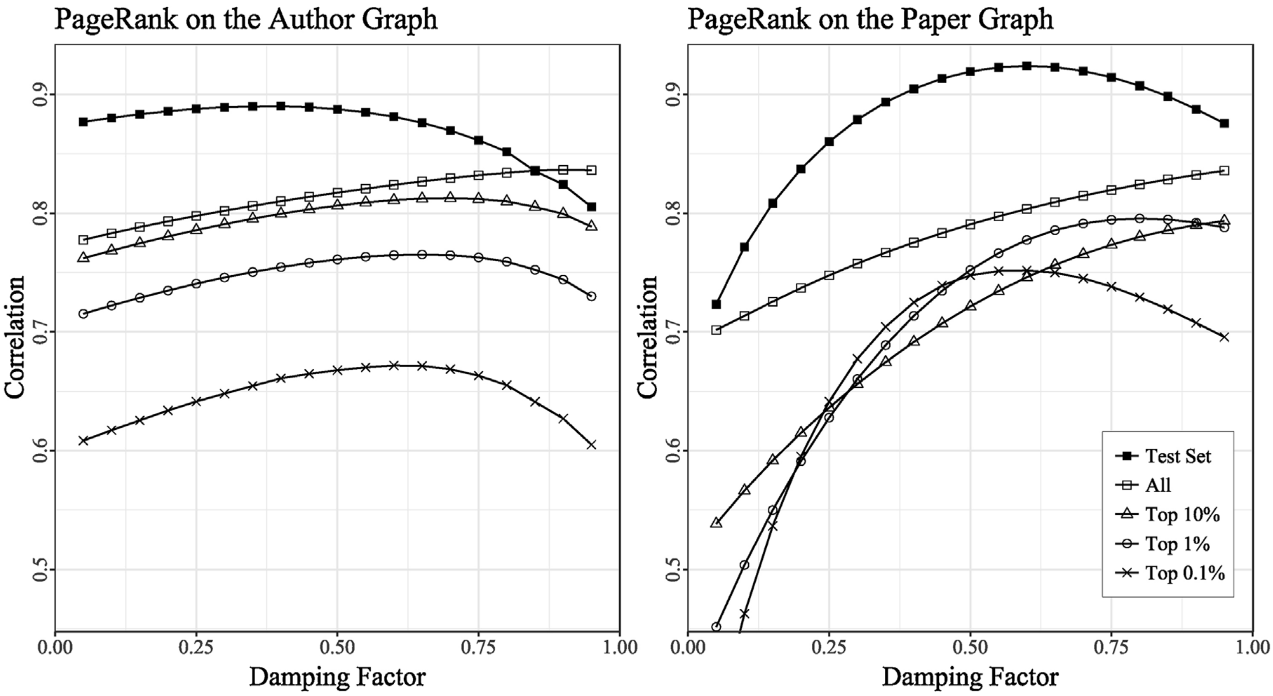


Fig. 7. Correlation between ranks according to citation counts and PageRank with varying damping factors based on the ACM database. The figure on the left shows the results of PageRank on the author citation graph, while the right figure shows the results of PageRank *DIV* on paper citation graph.

we found that either no personalisation (PR(none)) or per-paper journal influence values (PR(IF) and PR(AI)) yield the best results (see Table A.10). When the damping factor is optimised for the whole test data set on the MAG database, the entities are ranked highest when PageRank is personalised with the Eigenfactor values (PR(EF)). However, on the ACM database a better result is obtained when no personalisation is used.

Fiala et al., 2015 found that using PageRank-like impact metrics on the author graph yield no significant improvement over using citation counts. According to our results, we cannot confirm this result and instead show that using PageRank, either on the paper citation graph or on the author citation graph, significantly outperforms citation counts. In addition, their results showed that PageRank performs the best when the damping factor α is set to 0.5.

Nykl et al. (2015) found that the optimal α values are in the range from 0.55 to 0.85 when various PageRank personalisation strategies are used on the Web of Science publication data using ACM fellowships and two computer science awards as evaluation data. In contrast, Dunaiski et al. (2016) found that the author-level Eigenfactor algorithm performs the best in identifying high-impact authors when the damping factor is set to 0.92 for the ACM database and 0.84 for the MAS database. In this paper, we also found that the optimal α value is significantly higher and falls into the range of 0.85 and 0.95.

When PageRank is used to compute paper impact scores which are distributed among co-authors we also found that the optimal α values are high (0.95 and 0.9 for the ACM and MAG databases). Chen et al. (2007) used a damping factor of 0.5 based on the empirical observation that in their reference set of papers, about 42% of the papers referenced by a paper A have at least one reference directly to another paper that is also in the reference list of A. They argue that choosing a damping factor of 0.5 is more appropriate since this leads to an average citation path of length 2 in the PageRank model[2]. Walker, Xie, Yan, and Maslov (2007) showed that using a damping factor of 0.5 best predicts new citations to papers on two different publication data sets from the field of physics.

In the following discussion we try to shed some light into why we found such high values for the optimal damping factor. Fig. 7 shows the Spearman correlation coefficients between ranks by citation counts and PageRank with varied damping factor values based on the ACM database. The left figure shows the correlation between PageRank on the author graph and standard citation counts. The right figure shows the results when comparing PageRank on the paper graph (PageRank *DIV*) and Citations *DIV*. Different subsets of ranks are considered. For example, “Top 1%” shows the correlation when only the top 1% of authors (according to citation counts) are used. From these figures we can observe that for smaller subsets of highly cited authors, α moves towards the middle of the damping factor range. For the top 0.1% of authors, the highest correlation between citation counts and PageRank on the author graph (ρ = 0.672) and the paper graph (ρ = 0.752) is found when α = 0.6.

This also indicates that the optimal average citation path length to highly cited authors in the ACM database is approximately 2.5. This corroborates the results found by Ding et al. (2009) who, using the 108 most highly cited authors in the field of information retrieval, based on a subset of the WoS publication data, found that PageRank on the author citation graph

[2] The average path length is $\frac{1}{1-\alpha}$ in the PageRank model (Chen et al., 2007; Ding, Yan, Frazho, & Caverlee, 2009).

has the highest correlation to citation counts when α is set to 0.55. Similarly for the paper citation graph, Dunaiski (2014) found the closest correlation to citation counts when α varies between 0.65 and 0.75 when considering all paper entities in a citation network comprising computer science papers of the MAS database.

For unweighted paper-level citation graphs, PageRank's damping factor has a large influence on the ranking outcome due to the intrinsic time-directed nature of the underlying graph. For larger α values, higher average scores are assigned to older papers [p. 97] (Dunaiski, 2014). The influence of publication ages decreases for smaller α values and when α tends towards 0 the all paper scores are roughly the same (Chen et al., 2007). For author citation graphs the same properties are true when PageRank is not personalised. However, compared to citation counts, the influence that α has on the time-variation is smaller for PageRank on the author citation graph than on the paper citation graph.

Therefore, given a reference set of only highly cited authors or papers that are, on average, the same age as the publication database, we expect the average citation path length to be between 2 ($\alpha = 0.5$) and 4 ($\alpha = 0.75$) in the PageRank model.

The average publication year of the papers from researchers in our test data is 1999.87 which is in the older part of the ACM database (27th percentile). In comparison, the average publication year of all papers is 2003.29 (37th percentile). Moreover, not all of the researchers in out test data are highly cited. The same is true for the papers associated with the researchers in the test data. Therefore, it is not unexpected that the damping factor is relatively high for this test data set.

Due to these two observations, we would expect that PageRank on the paper graph performs better than citation counts for α larger than 0.6. Furthermore, for small α values the intrinsic normalisation impact of the damping factor will lead to poorer performance compared to citation counts. In the case of this data set and the ACM database, this cross-over point occurs at $\alpha = 0.4$. Interestingly, we observed that PageRank on the author graph always performs better than citation counts even for very small damping factors. We found that, even with $\alpha = 0.05$, PageRank still assign higher scores than citation counts to the oldest nodes in the author graph. It is unclear what the contributing factors are and is an interesting question for future investigations.

We believe that the reason for the high damping factors is a combination of a relatively old test data set in comparison to the average age of the publication databases and the fact that not all researchers in the test data set are highly cited. It should be highlighted that comparisons are difficult since α is highly dependent on the test data (specifically the age), the link structure of the underlying network, and the PageRank personalisation strategy used (Fiala & Tutoky, 2017).

6. Threats to validity

6.1. Internal validity

There are various uncontrolled factors that might have influenced the results of the evaluations. The main threat to validity would be the manual matching of the author names in the test data to their corresponding entities in the databases. On the ACM database, the matching could be validated since all ACM fellows' author entities are associated with an ACM fellowship badge and many of the awards are also associated with the authors. Therefore, this is mostly of concern for the MAG database matching. However, using the ACM fellows and LCA recipients, the obtained results on the MAG database agree with the results when evaluated on the ACM database. Moreover, as described in Section 4, the relatively high correlations of publication counts between Google Scholar, ACM and MAG databases indicate a relatively accurate matching on the MAG database.

The MAG and the ACM data sets cannot be seen as two distinct publication databases since the data contained in the ACM database is a subset contained in the MAG database, and therefore the ACM citation graph should be interpreted as a subgraph of the MAG citation graph. This 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. After matching all papers in the ACM database to the MAG database, we found that ACM papers have 6.44 fewer citations on average. However, the Pearson correlation between the citation counts of the matched papers is relatively high with a coefficient of 0.77 when considering all papers with one or more citations.

6.2. External validity

The set of awards and fellowships attributed to researchers that we use as test data are handed out to authors for their long-lasting, significant and innovative contributions to their fields and are based on judgments by peers. However, this is not perfect evaluation data since some researchers might not have been nominated by their peers. Furthermore, the decisions by the selection committees are subjective and take other aspects of impacts 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. We removed researchers from the test data that clearly received the awards or fellowships for non-academic achievements. Considering the LCA recipients, it should be noted that we treated all author awards equally but some prizes might be more prestigious than others.

The test data is biased towards the field of computer science which is a threat to the generalisability of the results to academic disciplines outside of computer science. The test data is also biased towards certain countries and generalisations to other countries should be made with caution. Lastly, the test data only comprises well-established researchers since finding evaluation data for young or promising, so called, “rising stars” is difficult to obtain. However, a few researchers in the test data have very few papers. We also noticed that the authors in the test data set have an above average number of single-authored papers which can influence the generalisability of the results in Section 5.4.

7. Conclusion

Of the indicators evaluated in this paper, PageRank is the best metric for ranking the well-established authors in our test data sets and outperforms the percentile-based metric R6, as well as the more traditional impact indicators such as citation counts and the *h*-index. We showed this by using three different test data sets consisting of researchers that have obtained scholarly fellowships or won prestigious research awards that signify continued and high impact in their fields of research.

We found that it is more important to personalise the PageRank algorithm appropriately on the paper level than deciding whether to include or exclude self-citations. In general, the best results were obtained when PageRank is personalised with papers’ corresponding journal impact values. However, on the author level, we found that author graph normalisation is more important than personalisation.

Self-citations play an important role for all metrics. We found that the improvements by only removing direct author self-citation on the author graph are not significant for most metrics. However, when removing all co-author self-citation the performances of all metrics are significantly improved. Lastly, we also found that the *PR*-index is better in identifying well-established researchers compared to the more commonly used *h*-index and *g*-index.

We evaluated different approaches of sharing paper credit among co-authors and found that evenly distributing a paper’s score between co-authors always yields the best results, irrespective of what paper impact indicator is used. However, it is important to note that we only show this for well-established researchers and that for researchers in other stages of their careers this might be different.

Two important findings have to be pointed out: (1) evenly sharing paper credit among co-authors is the best approach and does not require knowledge about the author positions and (2) computing PageRank on the author citation graph is computationally much more expensive than computing PageRank directly on the paper citation graph and does not significantly improve the results. Based on these two findings, we conclude that PageRank on the paper citation graph is the clear favourite for computing author impact scores of well-established researchers.

Author contributions

Conceived and designed the analysis: Marcel Dunaiski.
Collected the data: Marcel Dunaiski.
Contributed data or analysis tools: Marcel Dunaiski.
Performed the analysis: Marcel Dunaiski.
Wrote the paper: Marcel Dunaiski.
Supervision and reviewing: Willem Visser.
Supervision, reviewing and proof-reading: Jaco Geldenhuys.

Appendix A. Additional results and information

Table A.9
Comparison of the different personalisation strategies for PageRank used on the author citation graphs. The table shows the corresponding significance levels when comparing personalisation approaches (none, *h*-index, publication counts, citation counts). Each table cell shows the confidence levels to which the ranking of the row is significantly better than the ranking of the column. The columns in a table cell correspond to the three test data sets (left to right: ACM fellows, LCA recipients, INT authors). Similarly, the rows correspond to the three self-citation strategies (top to bottom: including all self-citations, omitting direct author self-citations, omitting all co-author self-citations). For each combination, four significance values are shown which correspond to four author graph normalisaion strategies (**One**, **N**, **OneDivN**, **Eigenfactor**.)

	None			*h*-index			Papers			Citations		
None	–											, ,
h-index	.		. , . *	–						,		,
	,		. , .							.		.
	,		. . .							. ,		,
Papers	. , . *		. , .	, ,			–			. . ,		
	, , .		, , .	, ,						. . .		,
	, , .		. , .	, ,						* . .		,
Citations	*	.	. * *	,		. *		,	. *	–		
	.	,	* *			, *			. *			
	,		* *			, .			, *			

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

Table A.10
Comparison of the different paper ranking methods used to compute impact scores for papers which are evenly shared among co-authors to compute author impact values. Each table cell shows the confidence levels to which the ranking produced by the method of the row is significantly better than the ranking produced by the metric in the column. For each table cell the values in the left and right columns correspond to the ACM and MAG databases. The rows correspond to the three test data sets (ACM fellows, LCA recipients, INT authors).

	3PR		EF		IF		AI		PR(3PR)		PR(EF)		PR(CC)		PR(AC)		PR(none)		PR(IF)		PR(AI)	
3PR	–					*** ***																
EF	*** ** ***		–			*** ***																
IF	*** ** ***		*** ** ***		–																	
AI	*** ** ***		*** ** ***		* * .	*** ***	–		*													
PR(3PR)	*** ** ***	*** ***	*** ** ***	*** ***	** *	*** *** ***	*	*** ***	–					.								
PR(EF)	*** ** ***	*** ***	*** ** ***	*** ***	* ** *	*** *** ***	* .	*** ***	*		–			.								
PR(CC)	*** ** ***	***	*** ** ***	***	*** ** ***	*** *** ***	* ** ***	,	** * **		* * **		–		.				*		.	
PR(AC)	*** ** ***	***	*** ** ***	***	*** ** ***	*** *** ***	* ** ***	***	** * **		* * *				–				,			
PR(none)	*** ** ***	*** ***	*** ** ***	*** ***	*** ** ***	*** *** ***	** ** **	*** ***	*** * **		** * *		.	.	* ,		–		.		,	
PR(IF)	*** ** ***	***	*** ** ***	***	*** ** ***	*** *** ***	** * ***	,	** . **		** . **		.		*				–			
PR(AI)	*** ** ***	. ***	*** ** ***	, ***	* ** *	*** *** ***	*** ** ***	. ,	*** * **		** * **		*		*						–	

Note: Significance levels α: *** 0.005 ** 0.01 * 0.05 . 0.1 , 0.15.

Table A.11
Summary descriptions of the various awards used for compiling the test data sets. The discipline labels are very broad and not official classifications. The column "Year Range" indicates the first and last year for which the award is used in the test data sets.

Award	Name	Discipline	Count	Year range
AAAI	ACM – AAAI Allen Newell Award	Computer Science	23	1994–2015
ACL	ACL Lifetime Achievement Award	Computer Science	15	2002–2016
ACM Fellows	–	Computer Science	1000	1994–2015
AICHE	Founders Award for Outstanding Contributions to the Field of Chemical Engineering	Engineering	120	1958–2016
APSA	James Madison Award	Political Science	13	1978–2014
ICCV 1	PAMI Azriel Rosenfeld Lifetime Achievement Award	Computer Science	5	2007–2015
ICCV 2	PAMI Distinguished Researcher Award	Computer Science	8	2007–2015
ICDM	IEEE ICDM Research Contributions Award	Computer Science	14	2001–2016
IJCAI	Award for Research Excellence	Computer Science	16	1985–2016
ISSI	Derek John de Solla Price Memorial Medal	Social Science	27	1984–2015
MOS	The Dantzig Prize	Mathematics	19	1982–2015
PLDI	Programming Languages Achievement Award	Computer Science	28	1997–2016
SIGACT	Knuth Prize	Computer Science	16	1996–2016
SIGAI	The ACM/SIGAI Autonomous Agents Research Award	Computer Science	18	2001–2017
SIGARCH	ACM SIGARCH Maurice Wilkes Award	Computer Science	19	1998–2016
SIGCHI	SIGCHI Lifetime Research Award	Computer Science	20	1998–2017
SIGCOMM	Lifetime Contribution Award	Computer Science	30	1989–2016
SIGGRAPH	The Computer Graphics Achievement Award	Computer Science	35	1983–2016
SIGIR	Gerard Salton Award	Computer Science	11	1983–2015
SIGKDD	SIGKDD Innovations Award	Computer Science	16	2000–2016
SIGMETRICS	Achievement Award	Computer Science	11	2003–2013
SIGMOBILE	Outstanding Contributions Award	Computer Science	16	1996–2015
SIGMOD 1	SIGMOD Edgar F. Codd Innovations Award	Computer Science	25	1992–2016
SIGMOD 2	SIGMOD Contributions Award	Computer Science	25	1992–2016
SIGOPS	Mark Weiser Award	Computer Science	17	2001–2016
SIGSIM	ACM SIGSIM Distinguished Contributions Award	Computer Science	8	2007–2016
SIGSOFT	ACM SIGSOFT Outstanding Research Award	Computer Science	26	1997–2016
USENIX	USENIX Lifetime Achievement Award	Computer Science	15	1997–2014

References

Abramo, G., & D'Angelo, C. A. (2015). Ranking research institutions by the number of highly-cited articles per scientist? *Journal of Informetrics*, *9*(4), 915–923.

ACM, Inc. (2015). *ACM Digital Library*. http://dl.acm.org/

ACM, Inc. (2017a]). *ACM Fellows*. https://awards.acm.org/fellows/nominations

ACM, Inc. (2017b]). *Association for Computing Machinery*. https://www.acm.org/

Ajiferuke, I., Burell, Q., & Tague, J. (1988). Collaborative coefficient: A single measure of the degree of collaboration in research? *Scientometrics*, *14*(5), 421–433.

Aksnes, D. W. (2003). A macro study of self-citation? *Scientometrics*, *56*(2), 235–246.

Assimakis, N., & Adam, M. (2010). A new author's productivity index: *P*-index? *Scientometrics*, *85*(2), 415–427.

Bergstrom, C. T., & West, J. D. (2008). *Eigenfactor Score and Article Influence Score: Detailed methods. Technical Report November*. University of California Santa Barbara.

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics? *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.

Clarivate Analytics. (2017). *Web of science*. https://www.webofknowledge.com

Ding, Y., Yan, E., Frazho, A., & Caverlee, J. (2009). PageRank for ranking authors in co-citation networks? *Journal of the American Society for Information Science and Technology*, *60*(11), 2229–2243.

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

Dunaiski, M., Geldenhuys, J., & Visser, W. (2018). How to evaluate rankings of academic entities using test data. *Journal of Informetrics* (Under review).

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, NY, USA: ACM.

Dunaiski, M., Visser, W., & Geldenhuys, J. (2016). Evaluating paper and author ranking algorithms using impact and contribution awards? *Journal of Informetrics*, *10*(2), 392–407.

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

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), 334–348.

Fiala, D., & Tutoky, G. (2017). PageRank-based prediction of award-winning researchers and the impact of citations? *Journal of Informetrics*, *11*(4), 1044–1068.

Gao, C., Wang, Z., Li, X., Zhang, Z., & Zeng, W. (2016). PR-index: Using the *h*-index and PageRank for determining true impact. *PLOS ONE*, *11*(9), e0161755.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*(4), 359–375.

Hirsch, J. E. (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.

Howard, G. S., Cole, D. A., & Maxwell, S. E. (1987). Research productivity in psychology based on publication in the journals of the American Psychological Association. *American Psychologist*, *42*(11), 975–986.

Hwang, W.-S., Chae, S.-M., Kim, S.-W., & 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, NY, USA: ACM.

Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents? *Journal of the American Society for Information Science and Technology*, *62*(7), 1370–1381.

Mariani, M. S., Medo, M., & Zhang, Y. C. (2016). Identification of milestone papers through time-balanced network centrality? *Journal of Informetrics*, *10*(4), 1207–1223.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research? *Scientometrics*, *36*(3), 343–362.

Microsoft. (2017a]). *Academic Knowledge API*. https://azure.microsoft.com/en-us/services/cognitive-services/academic-knowledge/

Microsoft. (2017b]). *Microsoft academic graph*. https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

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.

Phelan, T. (1999). A compendium of issues for citation analysis? *Scientometrics*, *45*(1), 117–136.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, *12*(5), 297–312.

Schloss Dagstuhl. (2017). *The DBLP Computer Science Bibliography*. , http://dblp.uni-trier.de/http://dblp.uni-trier.de/. http://dblp.uni-trier.de/

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

Thelwall, M. (2016). Interpreting correlations between citation counts and other indicators? *Scientometrics*, *108*(1), 337–347.

Trueba, F. J., & Guerrero, H. (2004). A robust formula to credit authors for their publications? *Scientometrics*, *60*(2), 181–204.

Voorhees, E. M., & Buckley, C. (2002). The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval, SIGIR '02*. pp. 316–323. New York, NY, USA: ACM.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(6), P06010.

Waltman, L., & Schreiber, M. (2013). On the calculation of percentile-based bibliometric indicators? *Journal of the American Society for Information Science and Technology*, *64*(2), 372–379.

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.