Scientometrics
DOI 10.1007/s11192-016-2168-y

# $C^3$-index: a PageRank based multi-faceted metric for authors' performance measurement

**Dinesh Pradhan[1] · Partha Sarathi Paul[1] · Umesh Maheswari[1] · Subrata Nandi[1] · Tanmoy Chakraborty[2]**

Received: 8 July 2016


**Abstract** Ranking scientific authors is an important but challenging task, mostly due to the dynamic nature of the evolving scientific publications. The basic indicators of an author's productivity and impact are still the number of publications and the citation count (leading to the popular metrics such as h-index, g-index etc.). H-index and its popular variants are mostly effective in ranking highly-cited authors, thus fail to resolve ties while ranking medium-cited and low-cited authors who are majority in number. Therefore, these metrics are inefficient to predict the ability of promising young researchers at the beginning of their career. In this paper, we propose $C^3$-index that combines the effect of citations and collaborations of an author in a systematic way using a weighted multi-layered network to rank authors. We conduct our experiments on a massive publication dataset of Computer Science and show that—(1) $C^3$-index is consistent over time, which is one of the fundamental characteristics of a ranking metric, (2) $C^3$-index is as efficient as h-index and its variants to rank highly-cited authors, (3) $C^3$-index can act as a conflict resolution metric to break ties in the ranking of medium-cited and low-cited authors, (4) $C^3$-index can also be used to predict future achievers at the early stage of their career.

✉ Tanmoy Chakraborty
tanchak@umiacs.umd.edu

Dinesh Pradhan
dineshkrp@gmail.com

Partha Sarathi Paul
mtc0113@gmail.com

Umesh Maheswari
umeshmaheswari7@gmail.com

Subrata Nandi
subrata.nandi@gmail.com

[1] Department of Computer Science and Engineering, National Institute of Technology, Durgapur, India

[2] Department of Computer Science, University of Maryland, College Park, MD, USA

Published online: 09 November 2016

**Keywords** $C^3$-index · Author ranking · Multi-faceted measure · Multilayered network

## Introduction

> "...which indices are preferred depends on the question that is asked. No single index provides an optimal metric of science, whether scaled at the level of the individual scientist, topic, field, journal, or discipline."
>
> John T. Cacioppo (2008)

How do we quantify the quality of science? The question is neither rhetorical nor an emotional one; it is very much relevant to promotion committees, funding agencies, national academies, politicians and so on, in order to recognize and acknowledge quality research and prominent researchers. Identifying high-quality research is necessary for the advancement of science, but measuring the quality of research is even more important in today's world when scientists in different research fields are increasingly competing with each other for different purposes such as receiving research grants, publishing papers in prestigious vanues (conferences/journals) etc. The widely accepted approach is to check the bibliographic record of a researcher—that is, the number and the impact of publications. Researchers with very different bibliographic credentials may have the same h-index (Bornmann and Daniel 2009). Kosmulski pointed out that h-index is more suitable for the assessment of mature scientists who have published at least 50 papers and have h-indexes of at least 10 (Kosmulski 2006).

Assessment of science is important for many different reasons. For researchers at an early stage of their careers, a metric of scientific work may provide significant feedback to their progress and their exact position in the scientific world. For the recruitment committees in universities/research institutes, such a metric may simplify the task of wading through bunch of applications to select a list of potential applicants for the interview. For university administrators, these metrics may help to judge researchers seeking promotion or tenure. For the departmental chairs in an Institute, these metrics may help suggesting annual raises and the allocation of scarce departmental resources. For scientific societies, these metrics may influence selecting award recipients. For research granting agencies, an assessment of scientific fields would help identifying areas of progress and vitality. For legislative bodies and boards of directors, a measure of science may provide a means of documenting performance, ensuring accountability, and evaluating the return on their research investment. Measures of science may have other applications such as identifying the structure of science, the impact of academic journals, influential fields of research in current time, and factors that may contribute to new discoveries (Cacioppo 2016).

Several studies have been conducted by formulating the scientific progress in terms of networks (such as citation network, coauthorship/collaboration network) (Pradhan et al. 2016a, b; Chakraborty et al. 2014a, b, 2015a). Studies on coauthorship networks focus on network topology and network statistical mechanics (Yan and Ding 2009). Although our research also deals with citation and collaboration networks, we take a different approach by studying micro-level network properties, with the aim of applying centrality measures such as PageRank for impact analysis (Yan and Ding 2009).

In citation analysis, the number of citations reflects the impact of a scientific publication. This measurement considers each citation equally — a citation coming from an obscure paper has the same weight as one from a ground-breaking, highly-cited work

(Maslov and Redner 2008). Pinski and Narin (1976) were the first to note the difference between popularity and prestige in the bibliometric area. They proposed using the eigenvector of a journal citation matrix (i.e., similar to PageRank) corresponding to the principal eigenvalue to represent journal prestige. Bollen et al. (2006) defined journal prestige and popularity, and developed a weighted PageRank algorithm to measure them. They defined 'popular' journals as those which are cited frequently by journals with little prestige, and 'prestigious' journals as those which are cited by highly prestigious journals. Their definitions are recursive. Recently, Ding and Cronin (2011) extended this approach to authors and applied weighted citation count to measure researcher's prestige in the field of information retrieval. They defined the popularity of a researcher as the number of times he/she is cited (endorsed) in total, and prestige as the number of times he/she is cited by highly cited papers. The main idea behind this prestige measure is to use simple citation count but to give more weights to highly-cited papers.

Since scholarly activities are often represented in the form of complex networks where authors, journals, and papers are connected via citing/being cited or coauthored, the network topology can significantly influence the impact of an author, journal, or paper. The recent developments of large-scale networks and the success of PageRank demonstrate the influence of the network topology on scholarly data analysis. PageRank or weighted PageRank have performed well in representing the prestige of journals (Bollen et al. 2006; Falagas et al. 2008); however relatively few researchers have applied this concept to authors. Some of the works that addressed the issues include (Fiala et al. 2008; Ding 2011; Radicchi et al. 2009; Życzkowski 2010) etc. These papers are built following the notion of Ding and Cronin (2011) and clearly address the issue why PageRank or weighted PageRank based algorithms could be applied to author citation networks for measuring the popularity and the prestige of scholars.

In this paper, we use citation networks of authors, publications and journals, constructed from a massive publication dataset related to Computer Science domain. Our aim is to find a measure with which one can rank the authors of scientific papers appropriately. Our proposed method includes the adoption of the PageRank algorithm, which can be considered as a measure of prestige, as well as a measure of significance.

Recently, through a bibliometric analysis of the entire Italian university population working in the hard sciences over the period 2001–2005, Abramo et al. (2011) attempted to answer some of the questions related to bibliographic research. The results show that the researchers with top performance with respect to their national colleagues are also those who collaborate more abroad; but that the reverse is not always true. Collaboration is a fundamental aspect of scientific research activity. The reasons for collaboration are many, however most can probably be attributed to a "pragmatic attitude to collaboration" (Melin 2000).

In our present work, we propose an author performance metric called $C^3$-index that ranks authors based on their received citations as well as their collaboration profile through a PageRank based strategy. The proposed index has moderate correlation (60%) with h-index. It is observed that one of the component scores (ACI-score) of the proposed index has very strong correlation (98%) with h-index, but the other two component scores (PCI-score and AAI-score) have significantly less correlation (40–50%). These observations suggest that the proposed index carries more information than h-index. We further observe that a significant fraction of authors having high AAI-score during the start of the time-frame (1998–2008) have achieved significantly high h-index during the end of the time-frame. We also notice that the authors who we find reaching a certain performance level in terms of their h-index values during the end of a given time-frame reach moderately high

performance level according to the proposed index at least 4–5 years in advance. This observation indicates the future prediction capability of the proposed index as well.

## Related work

To propose strategies for ranking authors, researchers from citation analysis and other domains largely use publications made by the corresponding authors and the citations received by those publications. The seminal work by Hirsch (2005) proposed *h-index* considering both the number of publications and citations in a balanced way. H-index and its variants gained wide acceptance in the research community because they are easy to compute, though critics pointed out their limitations (Costas and Bordons 2007; Waltman and van Eck 2012; Waltman et al. 2012).

Analysis of publication trajectory of authors from the faculty of psychology during their first 7 years of post-doctoral studies revealed that the rate of publication increased each year following completion of their doctorate program (Byrnes 2007). The largest rate of increase in publication counts of peer-reviewed journal articles was observed in the first 4 years rather than in the 2 years immediately before their tenure. Publication count prior to tenure does not tell the whole story, of course. In an investigation of gender differences in scientific productivity, Long (1992) found that women publish fewer articles than men during the first decade of their career, but this difference is reversed later in their careers. According to a search of the ISI database, John Ridley Stroop published only three papers during his career. The articles has been cited 3810 times, whereas other two papers received less than 1% of the citations of former paper (Cacioppo 2008). Total number of citations is necessary for the evaluation of one's scientific merit, but it misses the point that Stroop's scientific contributions to psychology were limited primarily to his efforts prior to the completion of his Ph.D. Any good metric for scientific qualities of a researcher should capture in its quantification such instantaneous rise and fall of an author during her research career, which in our opinion, no naive metric for scientific impact is capable of.

Many alternative proposals were made in the line of h-index to overcome those limitations, as well as to use the power of h-index—Hirsch (2010) himself proposed $\hbar$-index (pronounced as *hbar-index*) that considers multiple coauthors of a paper that h-index omits; Egghe et al. (2006) proposed *g-index*. Jin et al. (2007) proposed *AR-index*, another interesting metric whose value may decrease over time due to aging of citations. This somehow could downgrade researchers who 'rest on their laurels' for long.

H-index and its variants are elegant as well as have a concrete mathematical foundation (Redner 2010)—these are integral measures and have narrow bounds (order of hundreds).[1] On the other hand, the number of authors engaged in active research nowadays are in the range of millions. This infers, through *pigeon-hole principle* that every single h-index instance is associated with a very large number of authors. Again, there are very few authors having high h-index, and most of the authors lie in low h-index region. So, we may expect a power-law behavior in the distribution of h-index (and its variants such as g-index) across author spectrum, which is shown in Fig. 1a, c). Narrow resolution for the middle and bottom liners in case of citation-count based indices restricts the research community from any fine-grained analysis of the authors residing in that part of the spectrum, which is very much essential for predicting the future position; even a Nobel laureate has to start his/her career with very low h-index. As observed from Fig. 1b, d, a

[1] http://www.webometrics.info/en/node/58.

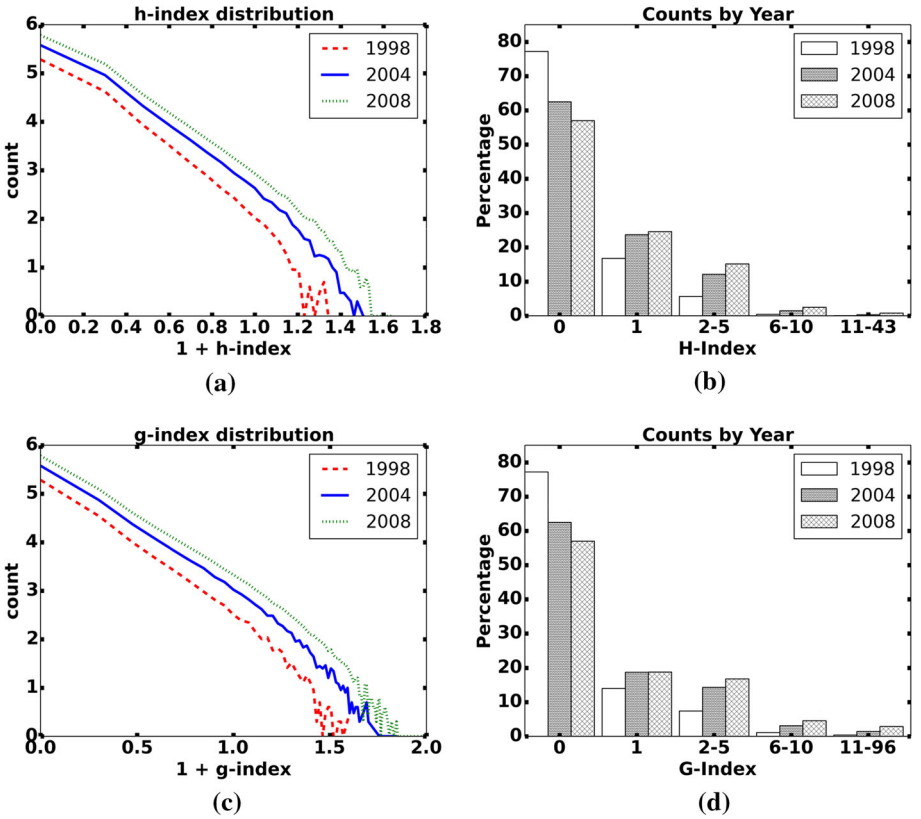



**Fig. 1** **a** Number of authors are plotted against different h-index values (plus one) in log-log scale for three different years: 1998, 2004 and 2008. **b** Percentage of authors as distributed across five different h-index bins for the year 1998 (*left bars*), distribution of the same set of authors in 2004 (*middle bars*), and in 2008 (*right bars*). The figure reflects that a limited fragment of authors attain high h-index over the years, but majority remains unimproved. **c**, **d** Show the similar plots against g-index. The near straight line nature of all the curves in **a** and **c** ensures power-law behavior of both h- and g-index. **b**, **d** Suggest that a small fragment of authors having low index values gradually improve over the years, whereas the majority remain unchanged. It is necessary to characterize as well as to predict, in well advance, the fragment of authors that have prospect of improvement

significant mass of the bottom liners gradually attain high h-index or g-index, whereas the remaining mass is nearly static over time.

H-index and its variants use only the citations received by individual papers of the concerned author for ranking the authors, though there are other features of an author that influence his/her career. Abramo et al. (2011) tested through an investigation on Italian university system that the authors who collaborate more at international level perform better than those who collaborate less; though the converse, they also observed, is not true in general. Credit sharing among coauthors of a multi-authored scientific paper is still an unresolved issue, though some attempts were made in this context. Trueba and Guerrero (2004) proposed a robust formula for the same, where the credit is shared among coauthors based on their relative position in the author name sequence in a paper. However such a distribution may not be fool-

proof, and may not be applicable to some domains,[2] where maintaining strict alphabetical name sequence is a common practice. Other concise study on the topic was undertaken by Xu et al. (2016) and Tscharntke et al. (2007), where they categorically discussed and analysed different proposed schemes, and pointed out the lack of any conclusive decision. A consensus about the credit sharing among coauthors is expected, because a promising star usually starts his/her career as a primary coauthor of an existing star; however the converse may not be true. Other influential factors may be the affiliation they have, the venues they choose for publication, the countries they belong to, and so on (Tahamtan et al. 2016).

As an alternative, some PageRank based schemes for ranking papers and authors have also been proposed. Chen et al. (2007) used Google PageRank method on the citation network formed by articles published in the Physical Review journals with the goal of measuring the importance of an individual scientific publication. They also pointed out that a choice of 0.5 as the damping factor suits better in the present context as compared to 0.85 in case of conventional webpage hyperlink network. Ma et al. (2008) claimed that a PageRank based representation might be a better indicator serving as a substitution of the number of citations for measuring the influence of a paper. Ding and Cronin (2011) differentiated the scholarly popularity versus the scholarly prestige of an author—the popularity of an author is the number of times the author is referenced by other papers, whereas the prestige is the number of times the author is cited by only highly-cited papers.

PageRank based methods for ranking authors use one or more from variety of features—citations from the peer colleagues, coauthorship with other researchers, co-citations for the papers/authors, and so on. Radicchi et al. (2009) proposed an author ranking algorithm based on diffusion of scientific credits using a proposed weighted author citation network. Barabási et al. (2002) studied the coauthorship among researchers as a social network and observed its dynamic behavior over time. Ortega (2014) observed that the structure of an author co-authorship network may reveal a good deal of information about research performance of a researcher through the analysis of the data from Microsoft Academic Search. Liu et al. (2005) studied the co-authorship network of the Digital Library (DL) research community as represented in the ADL, DL and JCDL conference series to reveal the structure of collaborations within the DL research community and the quantitative metrics for the concepts of status and influence. They built a weighted and directed network model to represent collaboration relationships, and proposed "AuthorRank", an alternative metric for ranking authors' prestige. Ding et al. (2009) observed the effects of different damping factors (ranging from 0.05 to 0.95) for author co-citation network, and noted that citation rank is close to PageRank for damping factor of 0.55. A detailed study on PageRank variants for ranking authors may be found in the work by Nykl et al. (2014).

*What would happen if one prefers to use more than one factor at a time in a PageRank based approach?* In a very recent work (Senanayake et al. 2015) and two of its preceding works (Senanayake et al. 2014a, b), Senanayeke et al. proposed *PageRank-Index* (*aka p-index* sometimes) that ranks authors by a PageRank-based approach using paper-paper citation and author-author coauthorship features at the same time. In their approach, the score for each paper is calculated using a PageRank-based approach; the score for each paper is distributed among all the coauthors of the paper in a weighted manner, where the weights are determined by the order of their authorship in the corresponding paper; finally, the PageRank shares obtained as above is summed and a percentile score is computed from the sum to get the final PageRank score. The approach is very effective, except a couple of points as follows— (a) the order of authorship may not be a standard measure of author

[2] https://en.wikipedia.org/wiki/Academic_authorship.

contribution in a paper, as mentioned earlier, (b) the proposed approach does not take advantage of the network property of author-author coauthorship network, and thus may miss some of the greater insights that the network could provide. In our work, we use network properties of author-author citation network and author-author coauthorship network for redistributing the PageRank based paper scores among the coauthors of the papers. To achieve this, we apply PageRank-based computation on these two layers (paper-paper citation network and author-author coauthorship network) as well.

In a student project at Stanford University, (Cui et al. 2010) proposed to represent the citation network as a multilayer network, which is the first attempt towards modeling multiple factors together for scholarly research impact metric. The technical report by Boccaletti et al. (2014) explains different aspects and applications of multilayer networks. It explicitly shows the areas where monoplex networks fail to capture the full detail of the scenario; whereas the multilayer networks may provide a better insight. Halu et al. (2013) proposed an idea of biased random walks to define the PageRank centrality measure on multiplex networks. De Domenico et al. (2015) claimed that calculating the centrality of nodes in component networks of the multilayer structure separately or aggregating the information to a single network leads to misleading results. They proposed to use tensorial formulation of multilayer networks to overcome the limitations.

There is a series of research that uses *heterogeneous networks*, which, in our observation, are very similar to multilayer networks. Zhou et al. (2007) used a heterogeneous network (similar to a two-layer network) for co-ranking authors and documents simultaneously. The co-ranking framework they adopted uses intra- and inter-class random walks to design a PageRank based strategy on heterogeneous networks. Expert finding with particular type of expertise for a given query is a non-trivial task, especially from a large-scale web systems, such as question answering and bibliography data, and is very much similar to the objective like author ranking. Deng et al. (2012) proposed a joint regularization framework to enhance expert retrieval by modeling heterogeneous networks as regularization constraints on top of document-centric model. Yan et al. (2011) used heterogeneous network similar to three-layer network model for ranking authors.

*Do PageRank based strategies reveal more information than simple citation-count based approaches?* Recently, Fiala et al. (2015) claimed that there is no evidence that PageRank based approaches certainly outperform simple citation-count based ranking approaches. The motivation of our work stems from this conclusion—we would like to design a PageRank-based ranking scheme that can provide additional information which may not be obtained from simple citation-count based strategies.

## Motivation

From the existing literature on author ranking strategies, one may observe that the following features are used for ranking authors: (a) paper-paper citation (Hirsch 2005; Egghe 2006; Hirsch 2010), (b) author-author citation (Ding and Cronin 2011), (c) author-author cocitation (Ding et al. 2009), (d) author-author coauthorship (Ortega 2014; Liu et al. 2005), (e) author-author collaboration[3] (Murthy and Lewis 2015).

[3] A supergraph of author-author coauthorship graph that takes into account social relationship between authors other than coauthorship: friends in the social media, Committee members of the same conference, editors of the same journal, members having same affiliation, etc. However, this feature is not frequently used due to the lack of suitable dataset.

Citations from other papers are possibly a well-accepted measure of the influence of a paper in the bibliographic domain. On the other hand, an author is best judged by the publications he/she made during his/his research life. A large pool of existing research follows this simple reasoning, and uses only *paper-paper citation network* to rank authors. However, not all papers of an author are of the same stature, and hence have different ranks. Now, combining these individual paper scores to a consistent author score might be challenging as well as debatable. An alternative solution might be to derive author rank solely from the author network; and possibly the first point to assume in this category is the one derived from paper-paper citations, viz, *author-author citation network*. Another, slightly less-intuitive is author-author cocitation networks, where two authors are connected if they cite the same set of papers. An intuitive justification might be that authors working on the same topic usually read and refer to the same set of papers.

One fundamental limitation associated with citation-based scoring technique is that it takes some time to gain attention after it is eventually published (Chakraborty et al. 2015b). Also the time required for a paper to be published after it is actually communicated to a venue is not small. Due to this factor, recent publications usually are misjudged if they are indexed only on the basis of citation count. The same limitation is observed in case of young authors, for whom major publications are quite recent and may not get enough attention in the early days.

To get rid of these limitations, other features such as author-author coauthorship or author-author collaboration were tried (Yan and Ding 2009). The reasons might be that high performers usually collaborate with other high-performers, either in case of coauthorship or in case of social collaboration. As an example, entry of a research student under an eminent researcher's supervision is quite a hurdle; so are the Technical Program Committee (TPC) members in a good conference, or to be an editor in a prestigious journal. We can thus assume that coauthors of an eminent researcher more or less have same calibre; and the same could be assumed for the TPC members of a good conference, or editors of a reputed journal. Hence, it is quite reasonable to exploit such social relationships to derive the influence of an author in the respective author network.

However, citation profile is a prominent feature to measure the influence of a paper, and hence removing it completely from the scope of study is unjustified. A better approach might be to devise a scoring strategy that considers all the above features mentioned earlier. However, we see that each individual feature leads to a complex network, directed or undirected, combining which leads to a multilayer complex network. It is quite evident that a PageRank like computation on a multilayer complex network is cost-inefficient. So the reduction of possible redundancy in the feature set, leading to the reduction of dimension of the underlying complex network, might save a huge amount of computation. If the resulting feature set (after removal of possible feature level redundancy) consists of only one feature, the resulting complex network would be a single layer network. Otherwise, we would try to find a minimal feature set that would lead to a multilayer complex network with least dimensions.

Through the reasoning so far, we find that author-author citation and author-author coauthorship are two indispensable features for any study for ranking authors. However, though an author-author citation network can be derived from a paper-paper citation network, the latter neither could replace the former, nor can be removed completely. The reason is that an author may not be judged completely without considering the quality of individual papers he/she has written. On the other hand, author-author citation relationship may not be avoided since prominent authors tend to write papers with their students or other (potentially) prominent authors. Note that we have excluded author-author social

collaboration relationship from our current study due to lack of data. Finally, we have excluded the author-author cocitation partly due to its less-intuitive nature, and mostly due to restricting computation by reducing the network dimension. On the summary, in our current study, we use three relationships—paper-paper citations, author-author citations and coauthor relationships among authors.

One particular issue that may seem confusing to the reader is that here we use paper-paper citation and author-author citation relationships simultaneously as features, where one can easily derive the latter from the former. An intuitive justification may be given as follows: a paper that is cited by a recent paper of an eminent researcher may receive equal credit to a citation by an unknown author; however by considering author-author citation we impost more credit to the former citation than the latter.

We now try to outline the proposed author ranking strategy and the underlying network model on which the proposed strategy would be applied.

## Network model and the outline of the proposed ranking strategy

In this paper, we propose a PageRank based multi-featured author indexing strategy called $C^3$-*index* (abbreviation of paper-paper **C**itations, author-author **C**itations and author-author **C**ollaborations) that may resolve generalized opinion among the majority class of low-profile authors. We shall see shortly that the proposed ranking scheme—is found to be consistent, effectively resolves the uncertainty among low-ranked authors, and may be used to predict future achievers in the early stage of their research career.

The $C^3$-index is developed on an underlying multi-layered citation-collaboration network model described in Fig. 2, where three layers from left to right correspond respectively to author-author citation network, author-author coauthorship network, and paper-paper citation network. The desired $C^3$-index score is obtained by the sum of three individual component scores obtained from three layers, scores being normalized

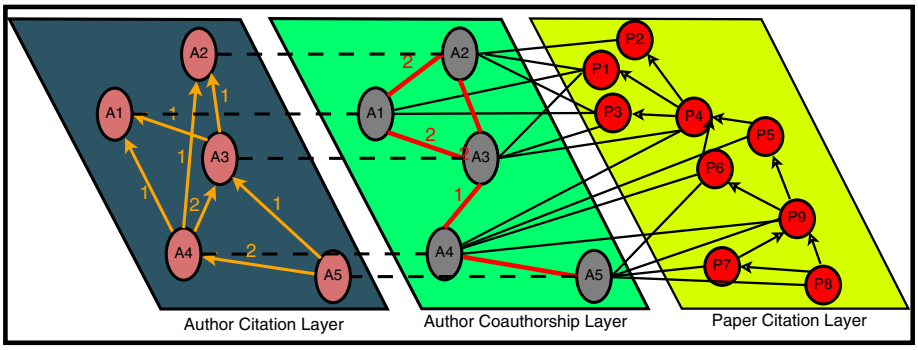



**Fig. 2** Three-layer network model used in $C^3$-index for ranking authors. Individual layers are: **a** Author citation layer- –a weighted directed network, where vertices are the authors, and weighted edges are drawn from vertex $A_j$ to $A_i$ if author $A_j$ cites the papers of author $A_i$, the weight of the edge being the number of papers of author $A_i$ being cited by author $A_j$; **b** Author coauthorship layer—a weighted undirected network where vertices are authors, and undirected weighted edges are given between authors who jointly published papers, the weight of the edge being the number of papers the pair coauthored; **c** Paper citation layer—a directed network where vertices are the papers, and edges are drawn from paper $P_j$ to paper $P_i$, if paper $P_j$ cites paper $P_i$. Lastly, there are inter-layer edges from author $A_i$ to paper $P_j$, if one of the authors in paper $P_j$ is $A_i$

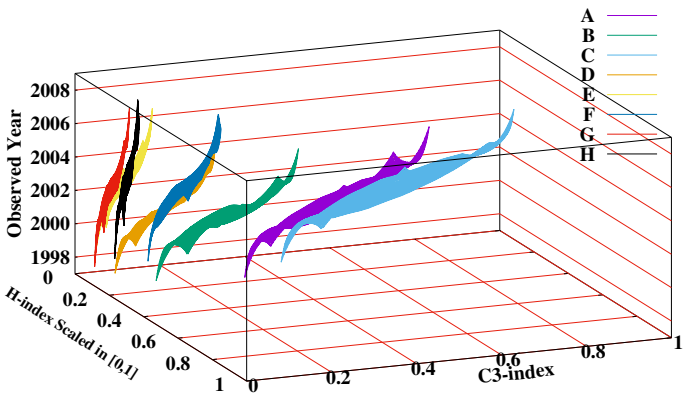



**Fig. 3** In the table, three component scores in $C^3$-index scoring strategy, viz. the Author Citation Index (ACI), Paper Citation Index (PCI), and Author coAuthorship Index (AAI) for eight selected authors are compared with their respective h- and g-index (the authors are selected from the results shown in Fig. 4, which will be discussed later). All the values of the metrics shown in the table are for the year 1998. A strong correlation may be observed between h-index, g-index and ACI component of the proposed $C^3$-index; but the same correlation is weak correlation with the other two components. These correlations suggest that the citation-based author ranking indices like h-index and g-index may nicely capture the effect of ACI component of the proposed $C^3$-index, but fail to capture the effects of the other two components. The 3D plot in the figure alongside shows the changes in h-index and $C^3$-index over the years for all the selected authors mentioned in the table during the year range 1998–2008. To maintain clarity of the figure, the h-index values are scaled within the range [0,1] by dividing actual h-index of the corresponding author by the observed maximum h-index value for an author in the dataset. Authors having higher $C^3$-index in 1998 show steeper growth both in h-index and in $C^3$-index as they progressed over the year, which may be an indication that $C^3$-index somehow captures the future success in advance

in such a way that the sum of scores of all the authors is unity. The component scores from individual layers are computed using PageRank based strategies on respective layers of the network. The strategy is elaborated in "Materials and methods" section. The table in Fig. 3 shows the $C^3$-index scores for eight selected authors along with individual score from each layer. For better visualization of the scores, the $C^3$-index score and its components are multiplied by the number of authors in dataset for the particular year, so that *the average $C^3$-index score for a particular year is always unity*. For the sake of comparison, we compute both h-index and g-index scores for the same authors in the table.

As we observe in Fig. 3, the surfaces corresponding to higher $C^3$-index scores in the beginning have steeper progress over the years both for h-index and $C^3$-index. This may be an indication that $C^3$-index can predict in advance the future success of the authors. This may be due to the AAI and PCI components in the $C^3$-index score (see "Materials and methods" section for detailed description), the former capturing the coauthor influence to the corresponding author, whereas the latter capturing the credit share of the authors due to their other coauthors. The rest of the paper is devoted to characterize $C^3$-index and to critically analyze whether it could be used for the purpose of predicting future prospect.

## Materials and methods

### Dataset collection, filtering and representation

We crawled a massive publication dataset related to Computer Science domain from Microsoft Academic Search (MAS), one of the largest archived datasets. Crawling of the Microsoft Academic Search started in October, 2015. The automated crawler initially used the ranklist given by MAS to obtain the list of paper IDs. The paper IDs were then used to fetch the metadata of the publications. We used Tor to distribute our crawling to different systems in order to avoid overloading a particular server with bursty traffic. We employed random exponential back-off time whenever the server or the connection returned some error and sent the request again. We followed the robot restrictions imposed by the servers to ensure efficient crawling of data from both client and server perspective. It took us around 6 weeks to completely crawl all the information related to the 7 million papers (Chakraborty et al. 2013, 2014b).

The crawled data had several inconsistencies that were removed through a series of steps. We filtered out all such papers that did not have the bibliographic attributes required for our study such as the unique index of the paper, the year of publication, the list of authors, the publication venue. We also removed few forward citations which pointed to the papers published after the publication of the source paper. Further, we considered only those papers published in between 1950 and 2012, that cite or are cited by at least one paper (i.e., we removed disconnected nodes with zero in-degree and zero out-degree). The filtered dataset contains around 6 million papers. Some of the references that pointed to papers absent in our dataset (i.e., dangling references) were also removed from the dataset. Some general information pertaining to the dataset are shown in Table 1.

### Network construction

From the filtered dataset, we prepare the multilayer network. The creation of the paper-paper citation network is easy—we consider the papers as the vertices, and connect vertex

**Table 1** General information of raw and filtered datasets

| | Raw | Filtered |
|---|---|---|
| Number of valid papers | 7,473,171 | 6,643,906 |
| Number of papers with no venue | 343,090 | – |
| Number of papers with no author | 45,551 | – |
| Number of papers with no publication year | 191,864 | – |
| Number of authors | 4,186,412 | 3,186,412 |
| Avg. number of papers per author | 5.18 | 5.04 |
| Avg. number of authors per paper | 2.49 | 2.67 |
| Number of unique publication venues | 6,143 | 5,938 |
| Number of paper-paper citation edges | – | 54,794,224 |
| Number of coauthorship edges | – | 10,837,179 |
| Sum of weights—coauthorship edges | – | 19,718,437 |
| Number of author-citation edges (excluding self loops) | – | 176,174,616 |
| Sum of weights—author citation edges | – | 371,572,974 |

$P_i$ to vertex $P_j$, if the paper corresponding to $P_i$ cites the paper corresponding to paper $P_j$. For preparing the other two layers, we need to extract the author information. To do that, the names of authors are extracted from the author lists for all the papers. To remove the ambiguity from author name, we use "RankMatch" algorithm proposed by Liu (2013). There are a couple of reasons behind adopting this algorithm. First of all, it is a completely unsupervised approach which is required in our study. In addition, the algorithm has been proved to be effective for the same types of scientific dataset. The algorithm first assigns a unique index ID to all the author names present in the dataset. Then it follows a two-step strategy–(1) For each indexing author ID, it tries to pull out all the authors whose names are possible variations of the indexing author name. To come up with the pool, it takes into account a number of cases where names can mutate or be disturbed. (2) In the second step, it trims the candidate pool based on authors' publication features. Examples of publication features include publication venues, years, and title words. These features turn out to be discriminative for identifying real duplicates from the candidate pool. Once the unique authors are extracted, they are given suitable Author Identifiers for further references. For each author in the Author List obtained here, we add one vertex each in the coauthorship layer as well as in the author-author citation layer. For each paper $P$ in the dataset, an edge between vertex $A_i$ and vertex $A_j$ is added in author-author coauthorship network, if both the authors corresponding to the IDs are in the author list of paper $P$. If there exists another paper $p'$ for which $A_i$ and $A_j$ coauthored together, the weight of the edge between vertex $A_i$ and vertex $A_j$ is incremented by one. For preparing the author-author citation network, we check if paper $P_i$ cites paper $P_j$, then all the authors of $P_i$ have links to all the authors of $P_j$. If some pair of authors already has links between them, its weight is incremented by one. Note that we also remove self-citation in the construction of the author-author citation network. We consider those citations as "self-citations" where at least one author is common in both citing and cited papers as defined in Carley et al. (2013). Once the network is created, the iterative modified PageRank algorithm discussed below are executed and the values for each vertex from different layers are collected.

### Measuring $C^3$-index

The proposed $C^3$-index is computed using a set of iterative formulas. The $C^3$-index of the $j$th author $A_j$ at iteration level $t$, denoted by $C_j^{3(t)}$, is obtained as:

$$C_j^{3(t)} = (1-\theta) + \theta \times (\mathrm{ACI}_j^{(t)} + \mathrm{AAI}_j^{(t)} + \mathrm{PCI}_j^{(t)})$$

In the above formula the terms $\mathrm{ACI}_j^{(t)}$ and $\mathrm{AAI}_j^{(t)}$, denote the scores of author $A_j$ in author-author citation network and the author-author coauthorship network, respectively, that are obtained using the following iterative formulas:

$$\mathrm{ACI}_j^{(t)} = (1-\theta) + \theta \times \sum_{A_k \in C(A_j)} \frac{\mathrm{ACI}_k^{(t-1)}}{\mathrm{outdeg}(A_k)}$$

$$\mathrm{AAI}_j^{(t)} = \sum_{A_k \in CA(A_j)} \frac{\mathrm{AAI}_k^{(t-1)}}{\deg(A_k)}$$

where $C(A_j)$ denote the set of authors who cited at least one paper of author $A_j$, $CA(A_j)$ denote the set of authors who coauthored with author $A_j$ in at least one paper, $\mathrm{outdeg}(A_k)$

denotes the sum of the degrees of the outgoing edges from node $A_k$ in the author-author citation layer of the network, $\deg(A_k)$ denotes the sum of the degrees of the edges incident on node $A_k$ in the author coauthorship layer, and $\theta$ is the *damping factor* for the PageRank based strategy. In our experiments, it is set to 0.5 following the suggestion made by Chen et al. (2007).

The third component in the formula, $\text{PCI}_j^{(t)}$ denotes the paper citation index score for author $A_j$ at the iteration level $t$ that are obtained from the paper citation layer of the network. It is the sum of the paper credits shared at that level for the publications made by author $A_j$ distributed uniformly (or some other rule) among all the coauthors of the paper using the formula:

$$\text{PCI}_j^{(t)} = \left(C_j^{3(t-1)}\right)^{\alpha} \times \sum_{P_k \in P(A_j)} \frac{\text{PQI}_k^{(t-1)}}{\sum_{A_l \in A(P_k)} \left(C_l^{3(t-1)}\right)^{\alpha}}$$

where $P(A_j)$ denote the set of papers published by the author $A_j$, $A(P_k)$ denote the set of authors for the paper $P_k$, and $\text{PQI}_k^{(t)}$ is a paper quality index score representing the credit of the paper that is obtained from the paper citation layer of the network using a PageRank based algorithm as follows:

$$\text{PQI}_i^{(t)} = (1-\theta) + \theta \times \sum_{P_k \in C(P_i)} \frac{\text{PQI}_k^{(t-1)}}{\text{outdeg}(P_k)}$$

where $C(P_i)$ denote the set of papers citing paper $P_i$, and $\text{outdeg}(P_k)$ denote the number of the outgoing edges from node $P_k$ of the paper citation layer. We use the same damping factor $\theta$ for all the PageRank formulas mentioned here.

As a final note, we represent $\text{PCI}_j$ as a generalized formula, where $\alpha$ is used as a *model parameter* to decide the way credit from an individual paper would be distributed among its authors. If it is set to 0, as is the case in the current experiments, then the credit will be distributed uniformly to all the coauthors. But for other values of $\alpha$, the credit will be distributed on the basis of their current $C^3$-index. If $\alpha$ is positive value, then authors having higher $C^3$-index would receive larger share of the credit, whereas if $\alpha$ is negative, the authors with lower $C^3$-index would receive larger share.

## Results

### $C^3$-index versus H-index

An immediate question would be how the ranking produced by h-index differs from the ranking obtained from $C^3$-index and its individual components. To verify this, we measure the Spearman Rank Correlation Coefficient between the pair-wise ranks (Table 2). The coefficient values suggest a strong correlation of h-index with ACI component, but relatively narrow correlation with the other two. This once again corroborate with our earlier observation in Fig. 3.

In Fig. 4, we show the correlation between $C^3$-index and h-index for all the authors in the dataset in a different manner. In all the sub-plots in Fig. 4, we plot the author scores obtained using $C^3$-indexing strategy for all the authors in the dataset against their

**Table 2** The Spearman rank correlation coefficient between h-index, $C^3$-index and its components

| Year | H-index versus $C^3$-index | H-index versus ACI | H-index versus PCI | H-index versus AAI |
|---|---|---|---|---|
| 1998 | 0.577136 | 0.989151 | 0.467660 | 0.401122 |
| 2004 | 0.604968 | 0.988483 | 0.517128 | 0.426008 |
| 2008 | 0.613174 | 0.988427 | 0.539801 | 0.437871 |

The values indicate that h-index is highly correlated with the ACI score, as compared to that for other two components, and hence with $C^3$-index as a whole. Thus we hypothesize that the information carried by $C^3$-index would be significantly different from that of h-index

respective h-index and g-index. The $C^3$-index as well as the h-index and g-index are calculated for a particular year by considering the publication entries in the dataset up to that particular year (i.e., by removing from the dataset the papers which are published after that year, the citations that are made after that year, and the authors who made their first publication after that year). The same procedure is followed for all the temporal studies made in this paper. In other words, as the year of study proceeds towards the current time, the data volume increases gradually in all respect.

In Fig. 4, the points close to the diagonal of each subplot represents those authors whose $C^3$-index values are perfectly correlated with h-index (g-index) strategy. However, we further observe that there are few authors with low h-index but high $C^3$-index (upper-left portion), and vice versa (lower-right portion). We selected some of these authors earlier and analyzed the profiles in Fig. 3.

## Temporal growth pattern

In Fig. 5 we study the year-wise transformation of performance indices (h-index as well as $C^3$-index) for four sets of authors selected from the authors in 1998. Figure 5a corresponds to the set of authors who have relatively low ACI-score, but high AAI-score in 1998. We select 31 authors from this category and show their growth of the said indices over the years. In Fig. 5b we plot similar results for 48 authors from the author pool who have low ACI-score and low AAI-score in 1998. While comparing the above two plots, we observe that the indices for most of the authors in Fig. 5a tend to end up with much higher values as compared to that for authors in Fig. 5b, in both the cases the lines start nearly from the same point. This perhaps hints upon a point that the ACI component of $C^3$-index has some kind of correlation with future performance behaviour of the concerned researcher. In Fig. 5c, d we plot similar growth curves for another two sets of authors, both having high ACI scores but different AAI scores. Here also we observe that the major portion of authors from the author set having higher AAI scores end up with higher performance indices.

## Capturing future performance through $C^3$-index

In Table 2 we already observed that h-index has strong correlation with the ACI component of $C^3$-index, but has weak correlation with the other two components. In Fig. 5, we intend to find whether that correlation behavior brings some meaningful insights about $C^3$-index. The figures suggest that authors having high AAI score show rapid growth over time than those the authors with low AAI score. From this, we hypothesize that the presence of this component in $C^3$-index may provide indication of future success, which h-index and

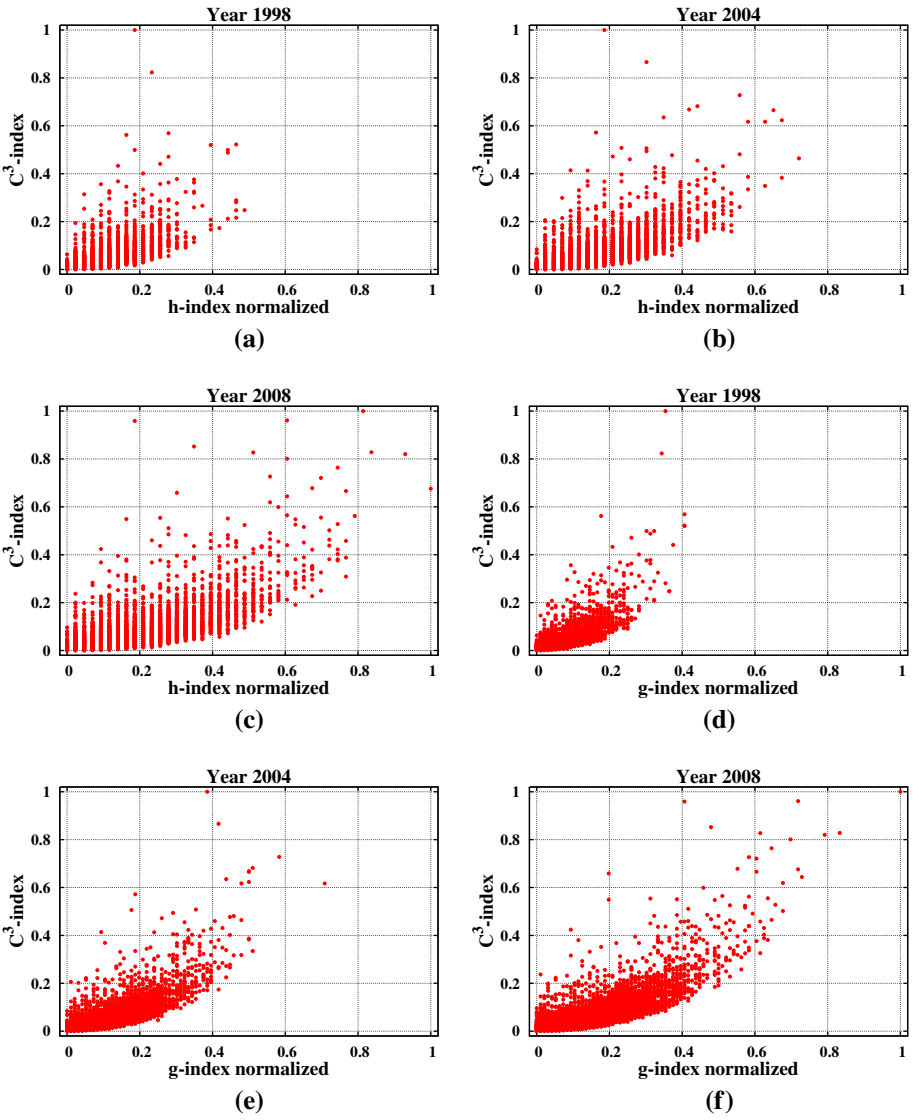



**Fig. 4** Scatter plots in the figure show distribution of $C^3$-index against h-index (*top panel*) as well as against g-index (*bottom panel*) for all the authors in the dataset during the years 1998, 2004 and 2008. Both h-index and g-index are scaled down within the range of 0–1 by dividing the actual index values by the highest value of the corresponding index in the time period of 1998–2008. In all the figures, we observe that the value of $C^3$-index for majority of the authors remains almost consistent with their respective h-index as well as with their g-index. However, we observe few inconsistent points mostly in upper-left portion of the plots, indicating those authors having low h-index (g-index), but high $C^3$-index. This is possibly an indication of low citation but high coauthorship credit for the corresponding authors. In Fig. 3, we selected some of authors having such inconsistencies and analyzed their behavior over the years

its variants perhaps lack. To validate this, we present multi-level pie-charts in Fig. 6 for a selected set of authors to show whether $C^3$-index is capable of predicting future success of authors in the early stage of their career.

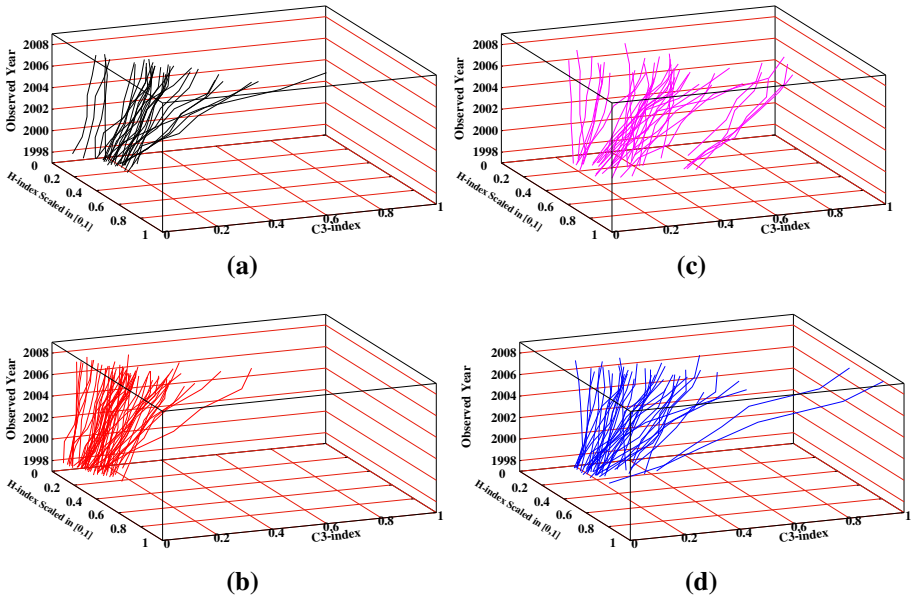
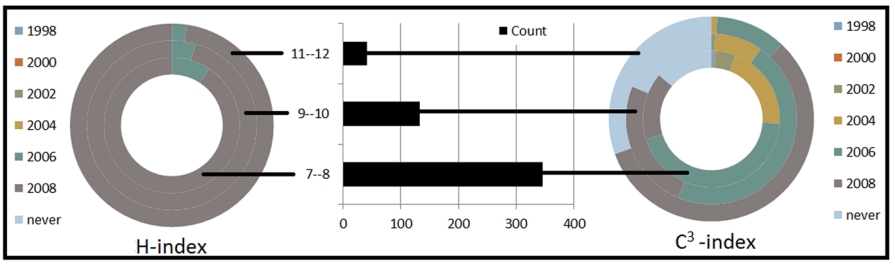


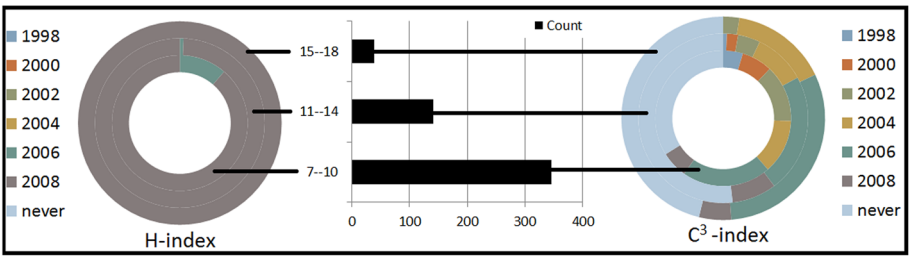

**Fig. 5** Proposed $C^3$-index has three components: ACI, PCI and AAI, respectively. We observed that h-index (and also g-index) has high correlation with ACI component, but has low correlation with the other two (Table 2). Here we select four sets of authors: **a** authors having ACI $\leq$ 20% of the $ACI_{max}$, AAI $\geq$ 80% of $AAI_{max}$, **b** authors having ACI $\leq$ 20% of the $ACI_{max}$, AAI $\leq$ 20% of $AAI_{max}$, **c** authors having ACI $\geq$ 80% of the $ACI_{max}$, AAI $\geq$ 80% of $AAI_{max}$, **d** authors having ACI $\geq$ 80% of the $ACI_{max}$, AAI $\leq$ 20% of $AAI_{max}$. The scores are selected on the basis of the year 1998. We plot 3D line curves for the corresponding authors in the respective sub-figures. In general, the figures suggest that the authors having high AAI-score improved more during the time period 1998–2008 than those having low AAI-scores. This suggests that the inclusion of AAI-score in the proposed $C^3$-index has brought future prediction capability in it

In Fig. 6a, the set of authors who had h-index of zero in 1998, but acquired moderate h-index (ranging from 7 to 12) in 2008 are selected. The bar plots in the middle show the number of such authors in three equally-divided h-index bins. The multi-level pie-chart in the left shows the gradual improvement of h-index as observed over time for the authors in each bin during the time span observed in 2-year separations. The multi-level pie-chart in the right points to the fraction of authors present in respective bins shown in the bar plot exceeding a chosen $C^3$-index bound in a given year. Three different bounds are chosen for three different bins, viz. 0.02 for 7–8 bin, 0.03 for 9–10 bin, and 0.04 for 11–12 bin. In left-hand pie-chart, we pin-point the fraction of authors that reached the next h-index bin in the respective year. We observe from the left-hand pie chart that no fraction of authors reach the next h-index bin prior to 2006. On the other hand, it is apparent from the right-hand pie-chart that significant fraction of authors reach the next bin level much earlier than the above, which suggests that $C^3$-index is able to capture the change much ahead of time than h-index. This in turn establishes the predictive power of $C^3$-index.

We are now interested to see whether above mentioned future-predictive behavior of $C^3$-index holds for authors present in other portion of the author spectrum. In Fig. 6b, a set of authors are selected whose h-index lay in the range of 4–7 in 1998. We may decently assume that such authors may be considered as medium-performers during the time when our observation begins. We observe that by 2008, the values of the selected authors' h-index lie in the range 7-18, which may indicate that some portion of the author gained high




**(a)**




**(b)**

**Fig. 6** **a** A set of authors is extracted from the dataset having zero h-index in 1998, but acquired moderate h-index (ranging 7–12) in 2008. The *bar plots in the middle* show the number of those authors in three equal-sized h-index bins. On the left-hand multi-level pie-chart, the three concentric rings correspond to the set of authors lie in the respective h-index bin associated. Each individual ring corresponds to a pie-chart that shows a distribution over the years of authors from author subset corresponding to the associated h-index bin meeting the h-index ceiling for that bin. The multi-level pie-chart on the right shows similar kind of distribution for the same sets of authors using their $C^3$-index over time. The $C^3$-index ranges for three concentric rings and are set to 0.02 - 0.029, 0.03 - 0.039, 0.04 - .... The *individual rings in pie-charts* represent similar author distribution over the years as on the *left-hand side figure*. The *pie-chart on the right side* suggests that in case of $C^3$-index, the change in the author score is visible much ahead of time than in the case of h-index, which indicates that the proposed strategy can capture the authors' future performance much ahead of time than h-index. **b** In order to verify whether the above observation is valid even for the cases of authors who already reached the level of medium/top rankers, a set of authors is extracted from the dataset having h-index ranging from 4 to 7 in 1998, but acquired moderate to high h-index (ranging 7–18) in 2008. The *bar plots in the middle* show the number of those authors in three equally-divided h-index bins as shown in the figure. We plot the same multi-level pie-chart pairs similar to **a**. In case of right-hand side chart, the $C^3$-index bins are set to be the following: 0.08 - 0.14, 0.141 - 0.17, 0.171 - .... The distributions of authors in both the multi-level pie-charts suggest that proposed $C^3$-index strategy can capture the future performance much ahead of time

visibility (i.e., gained high h-index) in 2008; whereas the rest fail to acquire enough visibility. The bar plot in the middle shows the number of those authors in three distinct bins similar to Fig. 6a. The multi-level pie-chart in the right pinpoints the fraction of authors lying in respective bins shown in the bar plot surpassing a chosen $C^3$-index bound in a given year. Three different bounds are chosen for three different bins, viz. 0.08 for 7–10 bin, 0.14 for 11–14 bin, and 0.17 for 15–18 bin. In left-hand pie-chart, we show the fraction of authors that reach the next h-index bin in the respective year. We observe from the diagram that major fraction of authors reach the next level after 2006, and only a small fraction reaches this level during 2006, and none does the same before 2006. On the other hand, for $C^3$-index, future stars (those falling in 15–18 bin), are capable of surpassing the predefined boundary set during 2004. For the others, it has been much earlier—a fraction,

although small, from 7–10 bin reaches this level even in 1998. This observation leads us to believe that proposed $C^3$-index has the capability of predicting future stars in advance.

## Discussion and future work

In our present work, we proposed a PageRank based multi-featured author ranking metric, *$C^3$-index*, that we expect would resolve some of the limitations that popular author ranking strategies such as h-index and its variants (g-index, $\hbar$-index, etc.) usually suffer from. One of the serious problems that we addressed here is the difficulty in ranking low-profile authors who are majority in number. The difficulty arises due to the fact that h-index and its variants produce integral scores spanning over a very low bounding range. The PageRank based strategy has been shown to overcome this problem.

The next issue we handled is the selection of features to devise the ranking strategy. We chose three features—the quality of citations received by the papers published by the concerned authors, the quality of citations received by the author from his/her peers, and the quality of coauthors he/she had worked with. All these features were collectively represented in the form of a multi-layer bibliographic network, which was further used for ranking the authors. There are three components in the expression for computing $C^3$-index, viz, ACI-score, PCI-score and AAI-score, each connected with one of the features mentioned above. We observed that popular author ranking indices like h-index and its variants have very high correlation with ACI-score component, but have significantly low correlation with the other two components. We may infer from this information that our proposed score carries more information about an author than h-index and its variants.

The third issue that has been addressed here is to find the relation of aforementioned components of $C^3$-index with the profile of the scientific authors. Temporal plots of $C^3$-index against h-index across the years reveal that the large fraction of authors having higher AAI-score component at a particular time attain larger values of h-index in future than authors having lower AAI-score. In other words, AAI component carries some indicator of future performance of an author within it. Interpreting differently, we may claim that $C^3$-index ranks an author not only on the basis of his/her present but also on his/her future prospect.

The fourth issue is to extend further the scope of future author performance prediction through the proposed strategy. We observed that $C^3$-index reveals the future outreach of a good fraction of the selected authors much earlier than the actual time they reached that milestone. This may indicate an additional scope of application for the proposed strategy than mere ranking of authors based on their present performance. We shall also check the results on other datasets from different domains such as Physics, Biology to strengthen our claims.

## References

Abramo, G., DAngelo, C. A., & Solazzi, M. (2011). Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics, 5*(1), 204–213. doi:10.1016/j.joi.2010.11.002.

Barabási, A. L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, *311*(3), 590–614.

Boccaletti, S., Bianconi, G., Criado, R., del Genio, C., Gmez-Gardees, J., Romance, M., et al. (2014). The structure and dynamics of multilayer networks. *Physics Reports*, *544*(1), 1–122. doi:10.1016/j.physrep.2014.07.001.

Bollen, J., Rodriquez, A. M., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, *69*(3), 669–687. doi:10.1007/s11192-006-0176-z.

Bornmann, L., & Daniel, H. D. (2009). The state of h index research. *EMBO Reports*, *10*(1), 2–6.

Byrnes, J. P. (2007). Publishing trends of psychology faculty during their pretenure years. *Psychological Science*, *18*(4), 283–286.

Cacioppo, J.T. (2008) Metrics of Science. *Association for Psychological Science, 21*(1). http://www.psychologicalscience.org/index.php/publications/observer/2008/january-08/metrics-of-science.html.

Cacioppo, J. T. (2016). *Social neuroscience*. Cambridge: MIT Press.

Carley, S., Porter, A. L., & Youtie, J. (2013). Toward a more precise definition of self-citation. *Scientometrics*, *94*(2), 777–780. doi:10.1007/s11192-012-0745-2.

Chakraborty, T., Ganguly, N., & Mukherjee, A. (2014a). Automatic classification of scientific groups as productive: An approach based on motif analysis. In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 130–137). doi:10.1109/ASONAM.2014.6921572.

Chakraborty, T., Ganguly, N., & Mukherjee, A. (2015a). An author is known by the context she keeps: Significance of network motifs in scientific collaborations. *Social Network Analysis and Mining*, *5*(1), 16:1–16:21. doi:10.1007/s13278-015-0255-3.

Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., & Mukherjee, A. (2015b). On the categorization of scientific citation profiles in computer science. *Communications of the ACM*, *58*(9), 82–90. doi:10.1145/2701412.

Chakraborty, T., Sikdar, S., Ganguly, N., & Mukherjee, A. (2014b). Citation interactions among computer science fields: A quantitative route to the rise and fall of scientific research. *Social Network Analysis and Mining*, *4*(1), 187. doi:10.1007/s13278-014-0187-3.

Chakraborty, T., Sikdar, S., Tammana, V., Ganguly, N., & Mukherjee, A. (2013) Computer science fields as ground-truth communities: Their impact, rise and fall. In *ASONAM* (pp. 426–433). Niagara Falls, Canada

Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8–15. doi:10.1016/j.joi.2006.06.001.

Costas, R., & Bordons, M. (2007). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, *1*(3), 193–203. doi:10.1016/j.joi.2007.02.001.

Cui, J., Wang, F., & Zhai, J. (2010). Citation Networks as a multi-layer graph: Link prediction and importance ranking. Stanford Student Project. http://snap.stanford.edu/class/cs224w-2010/proj2010/05_ProjectReport.pdf

De Domenico, M., Sole-Ribalta, A., Omodei, E., Gomez, S., Arenas, A. (2015). Ranking in interconnected multilayer networks reveals versatile nodes. *Nature Communications, 6*. doi:10.1038/ncomms7868

Deng, H., Han, J., Lyu, M.R., & King, I. (2012). Modeling and exploiting heterogeneous bibliographic networks for expertise ranking. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '12* (pp. 71–80). ACM, New York. doi:10.1145/2232817.2232833.

Ding, Y. (2011). Applying weighted pagerank to author citation networks. *Journal of the American Society for Information Science and Technology*, *62*(2), 236–245. doi:10.1002/asi.21452.

Ding, Y., & Cronin, B. (2011). Popular and/or prestigious? Measures of scholarly esteem. *Information Processing and Management*, *47*(1), 80–96. doi:10.1016/j.ipm.2010.01.002.

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. doi:10.1002/asi.21171.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152. doi:10.1007/s11192-006-0144-7.

Falagas, M. E., Kouranos, V. D., Arencibia-Jorge, R., & Karageorgopoulos, D. E. (2008). Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB Journal*, *22*(8), 2623–2628.

Fiala, D., Rousselot, F., & Ježek, K. (2008). PageRank for bibliographic networks. *Scientometrics*, *76*(1), 135–158. doi:10.1007/s11192-007-1908-4.

Fiala, D., Subelj, L., Zitnik, S., & Bajec, M. (2015). Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, *9*(2), 334–348. doi:10.1016/j.joi.2015.02.008.

Halu, A., Mondragn, R. J., Panzarasa, P., & Bianconi, G. (2013). Multiplex PageRank. *PLoS One, 8*(10), 1–10. doi:10.1371/journal.pone.0078293. http://dx.doi.org/10.1371%2Fjournal.pone.0078293.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences of the United States of America, 102*(46), 16569–16572. doi:10.1073/pnas.0507655102. http://www.pnas.org/content/102/46/16569.abstract.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, *85*(3), 741–754. doi:10.1007/s11192-010-0193-9.

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863. doi:10.1007/s11434-007-0145-9.

Kosmulski, M. (2006). A new Hirsch-type index saves time and works equally well as the original h-index. *ISSI Newsletter* (pp. 4–6).

Liu, J., Lei, K.H., Liu, J.Y., Wang, C., & Han, J. (2013). Ranking-based name matching for author disambiguation in bibliographic data. In *Proceedings of the 2013 KDD Cup 2013 Workshop, KDD Cup '13* (pp. 8:1–8:8). ACM, Chicago. doi:10.1145/2517288.2517296.

Liu, X., Bollen, J., Nelson, M. L., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information Processing and Management*, *41*(6), 1462–1480. doi:10.1016/j.ipm.2005.03.012.

Long, J. S. (1992). Measures of sex differences in scientific productivity. *Social Forces*, *71*(1), 159–178.

Ma, N., Guan, J., & Zhao, Y. (2008). Bringing PageRank to the citation analysis. *Information Processing and Management, 44*(2), 800–810. doi:10.1016/j.ipm.2007.06.006. Evaluating exploratory search systems digital libraries in the context of users broader activities.

Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of Neuroscience*, *28*(44), 11103–11105.

Melin, G. (2000). Pragmatism and self-organization: Research collaboration on the individual level. *Research Policy*, *29*(1), 31–40.

Murthy, D., & Lewis, J. P. (2015). Social media, collaboration, and scientific organizations. *American Behavioral Scientist, 59*(1), 149–171. doi:10.1177/0002764214540504. http://abs.sagepub.com/content/59/1/149.abstract.

Nykl, M., Jeek, K., Fiala, D., & Dostal, M. (2014). PageRank variants in the evaluation of citation networks. *Journal of Informetrics, 8*(3), 683–692. doi:10.1016/j.joi.2014.06.005.

Ortega, J. L. (2014). Influence of co-authorship networks in the research impact: Ego network analyses from microsoft academic search. *Journal of Informetrics*, *8*(3), 728–737. doi:10.1016/j.joi.2014.07.001.

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

Pradhan, D., Chakraborty, T., Pandit, S., & Nandi, S. (2016). On the discovery of success trajectories of authors. In *Proceedings of the 25th International Conference on World Wide Web (WWW)* (pp. 91–92). doi:10.1145/2872518.2889375.

Pradhan, D., Paul, P.S., Maheswari, U., Nandi, S., & Chakraborty, T. (2016). $C^3$-index: Revisiting author's performance measure. In *Proceedings of the 8th ACM Conference on Web Science, WebSci 2016*. Hannover, Germany, May 22–25, 2016 (pp. 318–319).

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5), 056–103.

Redner, S. (2010). On the meaning of the h-index. *Journal of Statistical Mechanics: Theory and Experiment*, *2010*(03), L03,005.

Senanayake, U., Piraveenan, M., & Zomaya, A. (2014). Ranking scientists from the field of quantum game theory using p-index. In *Foundations of Computational Intelligence (FOCI), 2014 IEEE Symposium on* (pp. 9–16). doi:10.1109/FOCI.2014.7007801

Senanayake, U., Piraveenan, M., & Zomaya, A. (2014). The p-index: Ranking scientists using network dynamics. *Procedia Computer Science, 29*, 465–477. doi:10.1016/j.procs.2014.05.042.

Senanayake, U., Piraveenan, M., & Zomaya, A. (2015). The PageRank-index: Going beyond citation counts in quantifying scientific impact of researchers. *PLoS One, 10*(8), 1–34. doi:10.1371/journal.pone.0134794. http://dx.doi.org/10.1371%2Fjournal.pone.0134794.

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: A comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225. doi:10.1007/s11192-016-1889-2.

Trueba, F. J., & Guerrero, H. (2004). A robust formula to credit authors for their publications. *Scientometrics*, *60*(2), 181–204. doi:10.1023/B:SCIE.0000027792.09362.3f.

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multiauthored publications. *PLOS Biology, 5*(1), 1–2. doi:10.1371/journal.pbio.0050018. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1769438/.

Waltman, L., Costas, R., & van Eck, N. J. (2012). Some limitations of the H index: A commentary on Ruscio and colleagues' analysis of bibliometric indices. *Measurement: Interdisciplinary Research and Perspectives*, *10*(3), 172–175. doi:10.1080/15366367.2012.716260.

Waltman, L., & van Eck, N. J. (2012). The Inconsistency of the H-index. *Journal of the American Society for Information Science and Technology, 63*(2), 406–415. doi:10.1002/asi.21678.

Xu, J., Ding, Y., Song, M., & Chambers, T. (2016). Author credit-assignment schemas: A comparison and analysis. *Journal of the Association for Information Science and Technology*, *67*(8), 1973–1989. doi:10.1002/asi.23495.

Yan, E., & Ding, Y. (2009). Applying centrality measures to impact analysis: A coauthorship network analysis. *Journal of the American Society for Information Science and Technology*, *60*(10), 2107–2118. doi:10.1002/asi.21128.

Yan, E., Ding, Y., & Sugimoto, C. R. (2011). P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, *62*(3), 467–477. doi:10.1002/asi.21461.

Zhou, D., Orshanskiy, S.A., Zha, H., & Giles, C.L. (2007). Co-ranking authors and documents in a heterogeneous network. In *Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE*, Omaha, Nebraska (pp. 739–744).

Życzkowski, K. (2010). Citation graph, weighted impact factors and performance indices. *Scientometrics*, *85*(1), 301–315.