Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation

Ilias Kanellos, Thanasis Vergoulis, Dimitris Sacharidis, Theodore Dalamagas, Yannis Vassiliou

Abstract—As the rate at which scientific work is published continues to increase, so does the need to discern high-impact publications. In recent years, there have been several approaches that seek to rank publications based on their expected citation-based impact. Despite this level of attention, this research area has not been systematically studied. Past literature often fails to distinguish between short-term impact, the current popularity of an article, and long-term impact, the overall influence of an article. Moreover, the evaluation methodologies applied vary widely and are inconsistent. In this work, we aim to fill these gaps, studying impact-based ranking theoretically and experimentally. First, we provide explicit definitions for short-term and long-term impact, and introduce the associated ranking problems. Then, we identify and classify the most important ideas employed by state-of-the-art methods. After studying various evaluation methodologies of the literature, we propose a specific benchmark framework that can help us better differentiate effectiveness across impact aspects. Using this framework we investigate: (1) the practical difference between ranking by short- and long-term impact, and (2) the effectiveness and efficiency of ranking methods in different settings. To avoid reporting results that are discipline-dependent, we perform our experiments using four datasets from different scientific disciplines.

Index Terms—Bibliometrics, Information Retrieval, Data Mining

1 Introduction

In the last decades, the growth rate of scientific publications, colloquially called *papers*, has been increasing — a trend that is expected to continue [1], [2]. This is not only due to the increase in the number of researchers worldwide [3], but also to the growing competition that pressures them to continuously produce publishable results, a trend known as "publish or perish" [4]. Meanwhile, large amounts of relevant data (e.g., manuscripts, citations, suppl. datasets) are being made publicly available thanks to open science initiatives (e.g., BOAI[1], cOAlation S[2], I4OC[3]).

Conventional query-dependent ranking mechanisms have been used to help researchers identify useful results and interesting insights from these data. These mechanisms rank documents based on their relevance to user-provided queries. However, popular query terms may result in thousands of relevant papers, with a large portion of them being of lower quality — "publish or perish" has been notoriously correlated with a significant drop in the average quality of scientific papers [5], [6]. It is thus apparent that such techniques should be combined with query-independent (also known as static-rank) methods that seek to rank papers based on their *impact*.[4]

While the impact of a paper "may be measured and understood in many different ways" [7] (e.g., downloads/views, social media attention), in this work we focus on citation-based impact, hereafter simply called *impact*. This only depends on the network formed from the citations that papers make. A multitude of researchers from various disciplines have proposed several paper ranking methods in recent years following this notion of scientific impact. Despite this level of attention, the field has not been systematically reviewed, let alone experimentally analyzed, for various reasons.

First of all, methods are often introduced by scientists of different disciplines, with each team ignoring the work of the other during evaluation. It has been additionally pointed out that there is "no comprehensive evaluation metric that is acknowledged by the academic community" for paper ranking methods [8]. To make things worse, in many cases different datasets and different experimental methodologies are used to evaluate the effectiveness of each method. As recently reported [9], developing benchmarks on unified and consistent scholarly datasets to enable the objective quantification of paper impact remains an important open issue in the field.

Moreover, hitherto literature overlooks the fact that the impact of a paper may be captured either in the short- or in the long-term. For example, an experienced researcher usually needs to search for *popular* papers, i.e., with high short-term impact, which are currently the focal point of the scientific community and which will gather many citations in the

- *I. Kanellos is a graduate student of NTU Athens, Greece and a Research Assistant at IMSI, 'Athena' RC, Greece. E-mail: ilias.kanellos@imis.athena-innovation.gr*
- *T. Vergoulis is a scientific associate at IMSI, 'Athena' RC, Greece.*
- *D. Sacharidis is with the E-Commerce Research Unit in TU Wien, Austria.*
- *T. Dalamagas is a researcher at IMSI, 'Athena' RC, Greece.*
- *Y. Vassiliou is a professor emeritus at NTU Athens, Greece.*

1. https://www.budapestopenaccessinitiative.org/
2. https://www.scienceeurope.org/coalition-s/
3. https://i4oc.org/

4. Indicative of the importance to the field of developing effective static-rank methods is the WSDM Cup 2016.

following years. On the other hand, another researcher may be interested in the *influential* papers, i.e., with high long-term impact, which have shaped the discipline she wants to delve in. Depending on user need, a different impact aspect might be preferable. As we later show, ranking methods designed with a particular impact aspect in mind may produce better ranking w.r.t. a different aspect. The settings in which a method is more advantageous and, more importantly, the particular features that make it superior in these settings, have not been adequately investigated in the past.

This work studies impact-based paper ranking methods, both theoretically and experimentally, based on their ability to capture the long-term (influence) and short-term (popularity) impact of papers. To the best of our knowledge, there is no systematic study examining both impact aspects, their differences, and how existing ranking methods define and exploit them. We discern only three related studies. The first [10] is a rather outdated experimental study of Web-inspired ranking methods, which inevitably ignores many popular existing approaches. Moreover, it does not evaluate against different impact aspects, and only includes experiments with papers from a single discipline (computer science), failing to provide generalizable conclusions. The second [9] provides a rather brief high-level categorization of the existing literature on paper impact measures, while the third [11] overlooks many relevant approaches. In addition, these last two studies fail to discern common ideas behind the methods proposed and elaborate on their differences. More importantly, though, no experimental comparison is provided.

The contributions of our work are the following:

- We posit that a great part of impact-based paper ranking methods capture one of two distinct paper impact aspects: *influence* and *popularity*, and we define the corresponding ranking problems (Section 2).
- We identify and classify the most important ideas employed in the literature, and further show how each method combines these ideas (Section 3).
- We study the various methodologies used in the experimental evaluation of paper ranking methods (Section 4).
- Based on our theoretical study of ranking methods, we carefully select a subset of them so that (1) they are representative of the entire field, and (2) we can draw conclusions about which specific idea is helpful for each impact aspect. Moreover, building on the classification of evaluation methodologies, we propose a specific benchmark framework that can help us differentiate between methods and impact aspects (Section 5).
- Our experimental evaluation uses four large datasets coming from different disciplines (two fields of physics, computer science, and life sciences) to avoid reporting results that are discipline-dependent. Our study has three goals. First, we want to investigate how distinct the notions of popularity and influence are in practice (Section 5.2). Second, we want to reveal which ranking methods and ideas perform best for influence (Section 5.3) and which for popularity (Section 5.4). Third, we want to examine how quickly iterative methods converge (Section 5.5).
- Last but not least, we provide scalable, open source implementations[5] for all experimentally validated methods in our study.

2 Problem Statement

We first present some background, and then proceed to introduce the two impact-based paper ranking problems we study.

2.1 Preliminaries

Citation Network. A *citation network* is a graph that has papers as nodes, and citations as edges. For a paper node, an outgoing (incoming) edge represents a reference to (from) another paper. A citation network is an evolving graph. While nodes' out-degrees remain constant, in-degrees increase over time when papers receive new references.

A citation network of N papers can be represented by the adjacency matrix $\boldsymbol{A}$, where the entry $A_{i,j} = 1$ if paper j cites paper i, i.e., there exists an edge $j \rightarrow i$ in the network, and 0 otherwise. We denote as t_i the *publication time* of paper i; this corresponds to the time when node i and its outgoing edges appear in the network.

In the following, we overview two node centrality metrics for citation networks.

Citation Count. The *citation count* of a paper i is the in-degree of its corresponding node, computed as $k_i = \sum_j A_{i,j}$. Note that we use k_i^{out} to denote the out-degree of paper i, i.e., the number of references paper i makes to other papers.

PageRank score. PageRank [12], [13] was introduced to measure the importance of a Web page. In the context of citation networks, the method simulates the behaviour of a "random researcher" that starts her work by reading a paper. Then, she either picks another paper to read from the reference list, or chooses any other paper in the network at random. The PageRank score of a paper i indicates the probability of a random researcher reading it, and satisfies:

$$s_i = \alpha \sum_j P_{i,j} s_j + (1-\alpha) v_i \tag{1}$$

where $\boldsymbol{P}$ is the network's transition matrix, $P_{i,j} = A_{i,j}/k_j^{out}$, and k_j^{out} is the out-degree of node j; $(1-\alpha)$ is the *random jump probability*, controlling how often the researcher chooses to read a random paper; and v_i is the *landing probability* of reaching node i after a random jump. Typically, each page is given a uniform landing probability of $v_i = 1/N$. Note that in the case of "dangling nodes", which contain no outgoing edges, i.e., nodes j with out-degree $k_j^{out} = 0$, the value of the transition matrix is undefined. To address this, the standard technique is to set $P_{i,j} = 1/N$, or to some other landing probability, whenever $k_j^{out} = 0$.

In the case of citation networks, researchers [14], [15] usually set $\alpha = 0.5$, instead of 0.85, which is typically used for the Web. The assumption is that a researcher moves by following references once, on average, before choosing a random paper to read. An in-depth analysis of PageRank and its mathematical properties can be found in [16].

5. https://github.com/diwis/PaperRanking

2.2 Rank by Influence and Popularity

Using the aforementioned node centrality metrics to capture the impact of a paper can introduce biases, e.g., against recent papers, and may render important papers harder to distinguish [14], [17], [18]. This is due to the inherent characteristics of citation networks, the most prominent of which is the delay between a paper's publication and its first citation, also known as "citation lag" [19], [20], [21], [22]. Thus, impact metrics should also account for the evolution of citation networks.

In this section, we formalize two citation network-based impact aspects that have been employed in the past, and which constitute the focus of our work. They are both based on node centrality metrics computed using future states of the citation network. Therefore, all methods that target these aspects, need to predict the ranking of papers according to an unknown future state. For what follows, let $\boldsymbol{A}(t)$ denote the snapshot of the adjacency matrix at time t, i.e., including only papers published until t and their citations. Further, let t_c denote current time.

Influence. The first impact aspect captures the long-term impact of a paper. The *influence* of a paper is its centrality metric at the citation network's ultimate state $\boldsymbol{A}(\infty)$ [23]. Based on the above, the following problem can be defined:

Problem 1. Given the state of the citation network at current time t_c, produce a ranking of papers that matches their ranking by influence, i.e., their expected centrality at state $\boldsymbol{A}(\infty)$.

Popularity. The second impact aspect captures the current, short-term impact of papers, which reflects the level of attention a paper enjoys *at present*, e.g., as researchers study it and base their work on it. The short-term impact can only be quantified by the citations a paper receives in the *near future*, i.e., once those citing papers are published. Exactly how long in the future one should wait for citations depends on the typical duration of the research cycle (preparation, peer-reviewing, and publishing) specific to each scientific discipline. Assuming this duration is T, the *popularity* of a paper is its centrality metric computed on the adjacency matrix $\boldsymbol{A}(t_c + T) - \boldsymbol{A}(t_c)$. Note that this matrix contains a non-zero entry only for citations made during the $[t_c, t_c+T]$ time interval.

If the centrality is citation count, popularity essentially counts the number of citations a paper receives in the near future. On the other hand, when popularity is defined by PageRank, it portrays the significance endowed to a paper by citation chains that occur during that time interval. Based on the above, the following problem can be defined:

Problem 2. Given the state of the citation network at current time t_c, produce a ranking of papers that matches their ranking by popularity, i.e., their expected centrality on adjacency matrix $\boldsymbol{A}(t_c + T) - \boldsymbol{A}(t_c)$, where T is a parameter.

3 Overview and Classification of Ranking Methods

In the following sections, we present a high-level overview of the various methods proposed for network-based ranking of scientific papers. Our goal is to identify the most important ideas that appear in the literature, and show how these ideas are adopted. We classify based on two main dimensions. The first concerns the type of time-awareness the methods employ. Here, we distinguish among no time-awareness, where simple adaptations of PageRank are proposed (Section 3.1); time-aware modifications in the adjacency matrix (Section 3.2); and time-awareness in the landing probability of a paper (Section 3.3). The second dimension is on the use of side information, where we discern between the exploitation of paper metadata (Section 3.4), and the analysis over multiple networks, e.g., paper-paper, paper-author, venue-paper (Section 3.5). Moreover, we cover methods that aggregate the results of multiple approaches that fall into one or several of the aforementioned categories (Section 3.6), and some that do not fit our classification (Section 3.7). Table 1 presents a classification summary.

3.1 Basic PageRank Variants

Here we refer to variations of PageRank's transition matrix $\boldsymbol{P}$, which utilize neither metadata nor time-based information. In each iteration, Non-Linear PageRank [24] computes a paper's score by summing the scores of its in-neighbors raised to the power of θ, and then taking its θ root, for some $0 < \theta < 1$. The effect is that the contribution from other important works is boosted, while the effect of citations from less important works is suppressed. SPRank [25], on the other hand, incorporates a similarity score in the transition matrix $\boldsymbol{P}$, which is calculated based on the overlap of common references between the citing and cited papers. In this way it simulates a focused researcher that tends to read similar papers to the one she currently reads. Another approach is SCEAS [10] that modifies PageRank's transition matrix by multiplying each entry by a quantity $q \in (0, 1)$, heavily decreasing the impact of longer paths to a particular paper's score. PrestigeRank [27] applies PageRank on a citation network that is artificially expanded by a virtual node. This node corresponds to all papers with citations given to, or received by papers not included in the dataset. The aim is to provide a fair ranking in cases where, for example, only few papers from a long reference list are included in the dataset, and thus are promoted by the citing paper more than they should. Finally, Focused PageRank [26] multiplies each item in the transition matrix with the percentage of each paper's citation count in a reference list, in an attempt to give advantage to the most cited papers.

3.2 Time-Aware Adjacency Matrix

The adjacency matrix $\boldsymbol{A}$ can include time quantities as weights on citation edges. There are three time quantities of interest, denoted as τ_{ij}, concerning a citation $j \to i$:

- *citation age*, or citing paper age, the elapsed time $t - t_j$ since the publication of the citing paper j,
- *citation gap*, the elapsed time $t_j - t_i$ from i's publication until its citation from j, and
- *cited paper age*, the elapsed time, $t - t_i$ since the publication of the cited paper i.

TABLE 1: Classification of Ranking Methods; bold font indicates the methods evaluated in the experimental section.

Method	Basic PR variants	Time Aware		Metadata		Multiple Networks	Ensemble	Other
		Network Matrix	Landing Probability	Venue	Author			
Non-Linear PageRank (NPR) [24]	✓							
SPR [25]	✓							
SCEAS [10]	✓							
Focused PageRank [26]	✓							
PrestigeRank [27]	✓							
Weighted Citation (WC) [28]		✓		✓				
Retained Adjacency Matrix (RAM) [29]		✓						
Timed PageRank [18], [30]		✓		✓	✓			
Effective Contagion Matrix (ECM) [29]		✓						
NewRank (NR) [8]		✓	✓					
NTUWeightedPR [31]		✓	✓	✓	✓			
EWPR [32]		✓		✓	✓		✓	
SARank [11]		✓		✓	✓		✓	
CiteRank (CR) [33]			✓					
FutureRank (FR) [34]			✓		✓	✓		
MR-Rank [35]		✓		✓		✓		
P-Rank [36]				✓	✓	✓		
YetRank (YR) [17]			✓	✓				
Wang et al. [37]			✓	✓	✓	✓		
COIRank. [38]			✓	✓	✓	✓		
PopRank [39]						✓		
MutualRank [40]						✓		
Tri-Rank [41]				✓	✓	✓		
NTUTriPartite (WSDM) [42]				✓	✓	✓	✓	
NTUEnsemble [43]		✓	✓	✓	✓	✓	✓	
bletchleypark [44]		✓		✓	✓		✓	
ALEF [45]					✓		✓	
S-RCR [46]								✓
Citation Wake [47]								✓
Age-Rescaled PR [48]								✓
Age- & Field- Rescaled PR [49]								✓
Bai et al. [50]								✓

The prevalent way to infuse time-awareness into the adjacency matrix is to weigh each non-zero entry $A_{i,j}$ by an exponentially decaying function of τ_{ij}:

$$A'_{i,j} = \kappa e^{-\gamma\tau_{ij}} A_{i,j},$$

where $\gamma > 0$ is the *decay rate*, and κ represents additional factors and/or a normalisation term.

If τ_{ij} is set to the citation age, recent citations gain greater importance. If τ_{ij} is set to the citation gap, citations received shortly after a paper is published gain greater importance. If τ_{ij} is set to the cited paper age, citations to more recently published papers gain greater importance. While it is possible to weigh citations based on any combination of these time quantities, we have not seen such an approach.

The effect of a time-aware adjacency matrix on degree centrality (citation count) is immediate: $\sum_j A'_{i,j}$ denotes a *weighted* citation count. This approach is taken by Weighted Citation [28] and MR-Rank [35] using citation gap, and by Retained Adjacency Matrix [29] using citation age.

In PageRank-like methods, the importance of a citation depends not only on the importance the citing paper carries, but also on a citation's time quantity. Timed PageRank [18], [30], and NewRank [8] adopt this idea using exponentially decayed citation, or cited paper age, and thus compute the score of paper i with a formula of the form:

$$s_i = \alpha \sum_j \kappa e^{-\gamma\tau_{ij}} P_{i,j} s_j + (1-\alpha) v_i \qquad (2)$$

Effective Contagion Matrix [29], is another time-aware method using citation age. However it is not based on PageRank, but on Katz centrality [51], i.e., compared to Equation 2 it uses the adjacency matrix $\boldsymbol{A}$ and does not calculate random jump probabilities. Other time-aware weights have also been proposed. For example, [31] uses a weight based on the ratio of the cited paper's number of citations divided by its age. Further, [32] and its extension called SARank [11] uses a citation age-based exponentially decaying weight, but only when the cited paper has reached its citation peak, i.e., the year receiving its largest number of citations.

3.3 Time-Aware Landing Probabilities

In PageRank and several PageRank-like methods, papers are assumed to all have an equal landing probability, but in several cases non-uniform probabilities are assigned. We denote as v_i the landing probability assigned to paper i. Past works assign landing probabilities that decay exponentially with the paper's age, i.e.,

$$v_i = \kappa e^{-\gamma(t-t_i)}.$$

This implies that newer papers have higher visibility than old ones. Note the contrast between the time quantities described in Section 3.2, which refer to *edges*, and the single time quantity, paper age, that concerns *nodes*.

We discern two ways in which landing probabilities can affect the network process. The first is in the probabilities of visiting a node after a random jump, similar in spirit to topic-sensitive [52] and personalised [53] PageRank. CiteRank [33], FutureRank [34], YetRank [17], and NewRank [8] compute the score of paper i with a formula of the form:

$$s_i = \alpha \sum_j P_{i,j} s_j + (1-\alpha)\kappa e^{-\gamma(t-t_i)}.$$

To be precise, NewRank also employs a time-aware transition matrix as per Section 3.2, while the process in FutureRank involves an additional authorship matrix for the paper-author network, as discussed in Section 3.5.

The second way is more subtle and concerns dangling nodes. Recall that the standard approach is to create artificial

edges to all other nodes assigning uniform transition probabilities. Instead, YetRank [17] assigns exponential decaying transition probabilities to dangling nodes as

$$P_{i,j} = \kappa e^{-\gamma(t-t_i)}, \text{ for all } i, j : k_j^{out} = 0.$$

3.4 Paper Metadata

Ranking methods may utilize paper metadata, such as author and venue information. Scores based on these metadata can be derived either through simple statistics calculated on paper scores (e.g., average paper scores for authors or venues), or from well-established measures such as the Journal Impact Factor [54], or the Eigenfactor [55].

The majority of approaches in this category incorporates paper metadata in PageRank-like models, to modify citation, or transition matrices and/or landing probabilities. Weighted Citation [28] modifies citation matrix $\boldsymbol{A}$ using weights based on the citing paper's publication journal. Thus, the method gives higher importance to citations made by papers published in high rank venues. YetRank [17] modifies PageRank's transition matrix $\boldsymbol{P}$ and landing probabilities $\boldsymbol{v}$. Particularly it uses journal impact factors to determine the likelihood of choosing a paper when starting a new random walk, or when moving from a dangling node to any other paper. This way, it simulates researchers that prefer choosing papers published in prestigious venues when beginning a random walk. NTUWeightedPR [31] modifies transition matrix $\boldsymbol{P}$ and landing probabilities $\boldsymbol{v}$. It uses weights calculated based on the cited paper's author, venue, and citation rate information, to simulate a "focused" researcher, who prefers following references to, or initiating a random walk from papers that are written by well-known authors, published in prestigious venues, and which receive many citations per year.

An alternative to the above approaches is presented in Timed PageRank [18], [30], which calculates the scores of recent papers, for which only limited citation information is currently available, solely based on metadata, while using a time-aware PageRank model for the rest. Particularly, scores for new papers are calculated based on averages (or similar statistics) of their authors' other paper scores, or based on average paper scores (or similar statistics) of other papers published in the same venue.

3.5 Multiple Networks

Ranking methods may also employ iterative processes on multiple interconnected networks (e.g., author-paper, venue-paper networks, etc.) in addition to the basic citation network. We can broadly discern two approaches: the first approach is based on mutual reinforcement, an idea originating from HITS [56]. In this approach ranking methods perform calculations on bipartite graphs where nodes on either side of the graph mutually reinforce each other (e.g., paper scores are used to calculate author scores and vice versa), in addition to calculations on homogeneous networks (e.g. paper-paper, author-author, etc). In the second approach, a single graph spanning heterogeneous nodes is used for all calculations.

The first of the aforementioned approaches is followed by FutureRank [34], P-Rank [36], MR-Rank [40], Wang et al. [37], COIRank [38] and Tri-Rank [41]. FutureRank combines PageRank on the citation graph with an author-paper score reinforcement calculation and an age-based factor. P-Rank uses both PageRank and mutual reinforcement calculations on author-paper and paper-venue graphs. MR-Rank uses a bipartite paper-venue graph, along with PageRank calculations on paper and venue graphs to rank papers and venues using linear combinations of their scores. Wang et al. use paper-venue and paper-author bipartite networks, as well as time-based weights to rank papers, authors, and venues. COIRank extends this model by modifying paper citation edges when their authors have previously collaborated, or when they work at the same institution. The goal is to reduce the effect of artificially boosted citation counts. Finally, Tri-Rank uses paper-author, paper-venue, and author-venue bipartite networks. It uses mutual reinforcement to iteratively calculate venue, paper, and author scores, alternatively using these bipartite networks and homogeneous networks of authors, papers, and venues. Additionally, Tri-Rank uses various weights applied on edges in these graphs, e.g., based on self citations between papers, or based on the order of authors in a paper's author list.

The second approach to using multiple networks is used by PopRank [39] and MutualRank [40]. PopRank simulates a "random object finder", an entity that conducts a random walk between connected web-pages and objects representing papers, authors, conferences and journals. The entity's transitions from one type of object to another, depend on learned probabilities called *popularity propagation factors*. MutualRank uses an adjacency matrix comprised of 3 inter- and 6 intra-network individual adjacency matrices. The intra-networks are weighted, directed graphs of papers, authors, and venues, while the inter-networks consist of edges between the aforementioned graphs (i.e., edges between papers and authors, etc). MutualRank ranks all of the aforementioned nodes, based on an eigenvector calculation on this aggregated adjacency matrix.

3.6 Ensemble Methods

Ensemble methods implement multiple ranking methods, and combine their results to come up with a single score per paper. The majority of the 2016 WSDM Cup[6] methods fall in this category. The goal of the Cup was to rank papers based on their "query-independent importance" using information from multiple interconnected networks [57].

NTUTriPartite [42], the winning solution of the cup, aggregates score propagations from various networks with a linear combination of each paper's in- and out-degree. This is done in an iterating fashion, using a predefined and fixed number of iterations.

NTUEnsemble [43], combines scores from the metadata-based version of PageRank proposed in NTUWeightedPR [31], with the cup's winning solution [42], and a method based on Wang et al [37]. In EWPR [32] paper scores are a combination of time-weighted PageRank scores calculated on a paper graph, time-weighted PageRank venue scores, calculated on a venue graph, and author scores, calculated as averages of their authored papers' PageRank

6. http://www.wsdm-conference.org/2016/wsdm-cup.html

scores. SARank [11] extends EWPR by including an additional score for papers based on exponentially weighted citation counts, as well as additional scores for authors and papers based on averages of this aforementioned citation count-based score. In ALEF [45], the authors use Article Level Eigenfactor (ALEF) (which is a PageRank-based method), to calculate paper scores. Based on these scores, they calculate author scores and then combine author and paper scores as a weighted sum. Finally, in bletchley-park [44], paper scores result as a linear combination of citation counts, PageRank scores, paper age, author, and venue scores, where author and venue scores are based on aggregations derived from their respective papers.

3.7 Other Methods

We discern a handful of methods that do not fall into any of the aforementioned categories. S-RCR [46] is a simplified version of the relative citation ratio RCR [58]. This method calculates the score for paper i, using its citation ratio (i.e., the number of citations it has received, divided by its age) and comparing it to that of all other papers in its "neighbourhood". A neighbourhood consists of all papers j that appear in any reference list together with paper i. Citation Wake [47] calculates the score of paper i, based on a normalized weighted sum of the cardinality of sets of papers j at shortest path distances l from i. Note, that each paper j is only counted in its shortest path distance from i, thus the method does not use all different paths that lead to paper i, but the set sizes of papers at increasing shortest paths. Age-Rescaled PageRank [48] and Age- and Field-Rescaled PageRank [49] calculate simple PageRank scores and then rescale them. In the case of Age-Rescaled PageRanks this is done based on the mean and standard deviation of the scores of n papers published before and after each paper in question. In the case of Age- and Field-Rescaled PageRank this rescaling is performed only based on papers published in the same field. Finally, Bai et al. [50] use an algorithm based on Quantum PageRank [59] where citation weights are a function of the geographical distance of the institutions of the citing and cited papers.

4 Overview and Classification of Evaluation Methodologies

The ranking methods presented in Section 3 often originate from various scientific communities, and may have different objectives, not always clearly defined. Moreover, each method is evaluated on diverse datasets, under varying assumptions and quantified with different metrics. The goal of this section is to clarify the objectives and classify the methodologies used in evaluation.

4.1 Evaluation of Ranking Effectiveness

Ground Truth Lists. One way to evaluate effectiveness is to use a *ground truth list* of papers that a method is expected to rank highly. Measures of agreement between lists are then used to quantify effectiveness. Typically, the ranking objective is to identify papers with long-term influence. Hence, the ground truth consists of award winning or selected important papers [8], [17], [24], [25], [36], [48], or papers co-authored by award winning authors [47]. This type of evaluation has inherent drawbacks, in that the lists may be partial, biased, or not available for certain disciplines.

User Judgements. Another way to evaluate ranking effectiveness is to use user judgments, as in [11], [39], and the 2016 WSDM Cup. In this evaluation type, trained experts give pairwise comparisons of papers and all rankings are evaluated against these human annotated data.

Held-Out Data. Another approach, followed by [11], [18], [24], [25], [29], [30], [33], [34], [35] is to assess how accurately a ranking method can predict the citation network-based impact aspects defined in Section 2.2. As these impact aspects are determined by a future state of the citation network, the evaluation approach involves *holding out* part of the dataset. Specifically, a timepoint t_c is chosen to represent the current time. This choice essentially creates a *present state* of the citation network captured by the adjacency matrix $\boldsymbol{A}(t_c)$, and a *future state* captured by $\boldsymbol{A}(t_c + T)$, where T represents a time horizon.

The evaluation proceeds in three steps. First, the *ground truth ranking* is produced, according to a desired impact aspect. In the second step, the ranking method produces a ranking of papers based solely on the present state of the citation network. In the third, the agreement of the method's produced ranking with the ground truth is quantified. Regarding the selection of the ground truth in the first step, work in [18], [29], [30], [33] considers popularity in terms of citation counts, a case we denote as P-CC; work in [34] considers popularity with respect to PageRank, which we denote as P-PR; and [29] considers influence based on PageRank, which we denote as I-PR. To the best of our knowledge, there is no work evaluating influence in terms of citation counts, which we denote as I-CC.

A last point to discuss concerns the choice of t_c. Recall that influence captures the long-term impact, defined by the expected centrality captured at the network's state infinite time units ahead in time. On the one hand, to accurately estimate influence, we need to consider a future state that is well ahead the present state, which means a small t_c value. On the other hand, to fairly evaluate a ranking method, we need to have a sufficiently large present state for it to learn from, which means a large t_c value. Clearly, there are conflicting requirements. In the case of popularity t_c is always selected to be close to the latest timestamp of the dataset. The implications of choosing t_c have not been investigated, where prior work typically sets t_c to a particular publication year, or so that the number of citations in present and future state have a particular ratio.

4.2 Other Evaluation Approaches

Descriptive Evaluation. There are some approaches that only indirectly assess ranking effectiveness. Exactly what is expected of a paper ranking method is however not apparent. In some cases, the top-ranked papers are provided along with a rationale of why the resulting ranking is valid or interesting [8], [14], [15], [18], [27], [30], [34], [47]. More often, the relationship of the produced ranked list with that of a baseline (citation count, PageRank, or other methods) is investigated. Typically, scatter plots [8], [14], [15], [28], [33], and rank correlation values [15], [25], [27], [35], [41] are used.

Another approach, taken in [14], [17], is the presentation of the average rank score per publication year of ranked papers. This type of evaluation can be useful to demonstrate that a method does not discriminate much against recently published papers.

Non-Effectiveness Evaluation. Some studies focus on evaluating aspects besides ranking effectiveness. For example, [33], [34] present measurements of computation time, and [25], [34], [35], [41] examine quickness of convergence. In [24], [25], methods are evaluated in terms of robustness against malicious behaviour (e.g., self citations). In [18], [30], [47], a different type of robustness is evaluated, namely the method's stability as its parameters are varied.

5 Evaluation

In this section, we first detail our evaluation methodology (Section 5.1), including the research questions we pose, then proceed to answer them (Sections 5.2–5.5), and finally discuss our findings (Section 5.6).

5.1 Evaluation Methodology

We start by stating our research goals, and then explain how we set to achieve them, describing the datasets used, the methods investigated, and the evaluation metrics used.

5.1.1 Research Questions

Our evaluation is centered around the following questions:

RQ1: How distinct are the notions of popularity and influence, and how distinct are their citation-count and PageRank flavors? In Section 5.2, we investigate the correlation between ground truth rankings according to I-CC, I-PR, P-CC, and P-PR. Our objective is to examine their differences and to verify that they capture different paper impact aspects.

RQ2: Which ranking methods perform best for each impact aspect? We evaluate the effectiveness of state-of-the-art methods in ranking papers based on their influence (Section 5.3) and popularity (Section 5.4).

RQ3: What is each method's convergence rate? Most methods run iteratively until a convergence criterion is satisfied. Comparing these methods based on their convergence rate can reveal trade-offs in terms of the number of iterations (and, hence, running time) required for them to produce an effective ranking. This can reveal which method should be preferred among those exhibiting comparable ranking effectiveness, in scenarios with time-constraints. Thus, we evaluate the convergence rate and the execution time of all the iterative methods (Section 5.5).[7]

5.1.2 Datasets

For our experiments, we use four datasets:

- *hep-th*[8] of about $30,000$ papers on high energy physics, from arXiv's archive published from 1992 to 2003;
- *APS*[9] of about half a million papers from American Physical Society journals published from 1893 to 2014;
- *PMC*[10] of about 1.12 million life sciences open access papers published from 1896 to 2016;
- *DBLP*[11] of about 3 million papers recorded by DBLP, published from 1936 to 2018 [60].

We note that the first two datasets have been widely used in the literature (e.g., [14], [24], [29], [33], [34]), while the last two are representative of real large collections of papers (with over a million nodes each), from two prolific scientific domains, life sciences and computer science.

5.1.3 Ranking Methods

We study the following paper ranking methods, chosen so as to cover all classes presented in Section 3.

PageRank (PR). This is the algorithm presented in [13].

Non-Linear PageRank (NPR). This is the basic PageRank variant introduced in [24].

CiteRank (CR). This PageRank variant uses time-aware landing probabilities [33].

FutureRank (FR). This PageRank variant uses time-aware landing probabilities and multiple networks [34]. The key idea is to distribute author scores to their authored papers, and paper scores to their respective authors, an approach inspired from HITS [56].

Retained Adjacency Matrix (RAM). This citation count variant uses a citation age-weighted adjacency matrix [29].

Effective Contagion Matrix (ECM). This Katz centrality-based method operates over a citation age-weighted adjacency matrix [29].

NewRank (NR). This PageRank variant uses a weighted citation matrix, with weights based on the age of cited papers, as well as time-aware landing probabilities [8].

YetRank (YR). This PageRank variant employs landing probabilities that are time-aware and also depend on journal impact factors, computed over the past five-years [17].

NTUTriPartite (WSDM). This ensemble method is not time-aware and computes a score for a paper depending on its authors, venues, citing papers, and in- and out- degrees, using a fixed set of iterations [42].

Weighted Citation (WC). This citation count variant [28] uses a citation matrix weighted by citation gap and a function of the journal Eigenfactor [55], which is an indicator comparable to the journal impact factor. In our implementation, we have used Journal Impact Factor instead, computed over the past five years.

Note that we test the last three methods (YR, WSDM, WC) only on the PMC and DBLP datasets, which contain information about the journal of each paper, which these methods require. For each method, we experimented with the various configurations suggested in the method's original paper, but only present results for the one that performed best in our experiments.

All methods are implemented in Python 2.7 and are freely available through a GNU GPL licence.[12] All experiments were executed on a cluster of 10 VMs (of 4 cores and 8GBs RAM) provided by okeanos Cloud service [61].

7. Note that this evaluation excludes citation count-based methods, as well as those running on a fixed number of iterations.
8. http://www.cs.cornell.edu/projects/kddcup/datasets.html
9. http://journals.aps.org/about
10. ftp://ftp.ncbi.nlm.nih.gov/pub/pmc
11. https://aminer.org/citation
12. https://github.com/diwis/PaperRanking

TABLE 2: Pairwise correlations (Spearman's ρ) of ground truth rankings for different future/present ratios η, per dataset.

hep-th	$\eta = 1.2$			$\eta = 1.4$			$\eta = 1.6$			$\eta = 1.8$			$\eta = 2$		
	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR
P-CC	0.690	0.649	0.376	0.747	0.747	0.453	0.775	0.806	0.520	0.790	0.846	0.560	0.794	0.861	0.596
P-PR		0.381	0.568		0.529	0.678		0.609	0.743		0.662	0.786		0.689	0.840
I-CC			0.840			0.823			0.820			0.814			0.817

APS	$\eta = 1.2$			$\eta = 1.4$			$\eta = 1.6$			$\eta = 1.8$			$\eta = 2$		
	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR
P-CC	0.834	0.541	0.387	0.869	0.650	0.474	0.883	0.715	0.527	0.889	0.761	0.567	0.884	0.780	0.599
P-PR		0.410	0.594		0.547	0.665		0.626	0.712		0.680	0.747		0.715	0.788
I-CC			0.904			0.895			0.887			0.880			0.876

PMC	$\eta = 1.2$			$\eta = 1.4$			$\eta = 1.6$			$\eta = 1.8$			$\eta = 2$		
	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR
P-CC	0.513	0.596	0.390	0.602	0.728	0.487	0.648	0.800	0.548	0.681	0.850	0.584	0.689	0.855	0.620
P-PR		0.223	0.623		0.392	0.769		0.486	0.866		0.562	0.840		0.607	0.934
I-CC			0.871			0.852			0.844			0.812			0.822

DBLP	$\eta = 1.2$			$\eta = 1.4$			$\eta = 1.6$			$\eta = 1.8$			$\eta = 2$		
	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR	P-PR	I-CC	I-PR
P-CC	0.757	0.662	0.517	0.803	0.757	0.591	0.825	0.810	0.635	0.837	0.844	0.672	0.835	0.854	0.693
P-PR		0.436	0.761		0.566	0.829		0.641	0.866		0.689	0.908		0.723	0.924
I-CC			0.939			0.927			0.918			0.916			0.910

All iterative methods were run with a convergence error set to 10^{-12}, to ensure that the final rankings produced are not subject to change if additional iterations would be performed.

5.1.4 Evaluation Metrics

To investigate our research questions, we conduct experiments based on the hold-out evaluation approach. Specifically, for each dataset, we fix the value of t_c so that the present state, $\boldsymbol{A}(t_c)$, contains half of the total papers in the dataset. To define the future state, $\boldsymbol{A}(t_c + T)$, we select T so that the *future/present ratio* η, measuring the ratio between the number of papers that appear in the future state and in the present state, takes values among $\{1.2, 1.4, 1.6, 1.8, 2.0\}$, with 1.6 being the default setting. Then, the ground truth is constructed from the future and present states, depending on the impact aspect considered, as detailed in Section 4.1.

For the first research question, we measure the correlation between pairs of ground truth rankings, using Spearman's ρ and Kendall's τ. For the second research question, we measure the correlation between the ranking of a method and a ground truth ranking using Spearman's ρ and Kendall's τ, and the method's ranking accuracy with respect to the ground truth using top-k precision and nDCG.

Spearman's ρ is based on the L_1 distance of all ranked items in the two lists [62]. *Kendall's* τ is computed based on the number of concordantly ordered pairs of items between the two lists [63]. *Top-k precision* calculates the percentage of common items among the top-k ranked items in each list. The discounted cumulative gain (DCG) at rank k of a paper is computed as $\text{DCG@}k = \sum_{i=1}^{k} \frac{rel(i)}{\log_2(i+1)}$, where $rel(i)$ is the ground truth score (I-CC, I-PR, P-CC, or P-PR) of the paper that appears at the i-th position on the method's ranking. The *normalized DCG at rank k* (nDCG@k) is the paper's DCG divided by the ideal DCG, achieved when the method's ranking coincides with the ground truth. The last two metrics are computed at a rank k, which takes values among $\{5, 10, 50, 100, 500\}$, with 50 being the default.

5.2 Comparison of Impact Aspects

The first research question investigates the relationship between the impact aspects. Specifically, we compute correlations among the four ground truth rankings, with respect to influence, either based on citation count (I-CC) or PageRank (I-PR), and with respect to popularity, based on citation count (P-CC) or PageRank (P-PR). Table 2 presents the correlation between pairs of rankings in terms of Spearman's ρ, as we vary the future/present ratio η; recall that $\eta = 1.2$, for instance, means that there are 20% more papers in the future state than in the present. Kendall's τ correlation shows similar trends and is, thus, omitted.

A general observation is that all ground truths are at least weakly correlated ($\rho > 0.2$) to each other, and in many cases very strongly ($\rho > 0.8$).[13] This implies a varying level of agreement among the impact aspects, meaning that the various impact aspects indeed capture different semantics.

An important observation is that the two flavors of impact aspect, citation count and PageRank, are correlated, strongly ($\rho > 0.6$) for popularity, and very strongly ($\rho > 0.8$) for influence. The latter may justify why past work has only considered one flavor of influence, namely I-PR. On the other hand, correlations *across* impact aspects exist, but are weaker. Naturally, the correlation across aspects is stronger when the same flavor is compared; e.g., 0.649 between P-CC, I-CC vs. 0.376 between P-CC, I-PR, for hep-th and $\eta = 1.2$.

Another trend is the increasing correlation between popularity and influence as the future/present ratio η becomes larger. Increasing η means that the future state $\boldsymbol{A}(t_c + T)$ grows with respect to the present state $\boldsymbol{A}(t_c)$, which is fixed. Thus, citations and citation chains in $\boldsymbol{A}(t_c + T) - \boldsymbol{A}(t_c)$ tend to become more similar to those in $\boldsymbol{A}(t_c + T)$, making centrality computed on the former graph closer to that computed on the latter.

Overall, we conclude that influence and popularity capture distinct paper impact semantics, which are however

13. We use the interpretation of correlation by Evans [64].

to some degree correlated[14]. The two flavors, citation count and PageRank produce relatively distinct rankings in terms of popularity, especially when η is small, which is the preferred setting for popularity. On the other hand, the two flavors produce highly similar rankings in terms of influence, particularly for large η values, which is the preferred mode for influence.

5.3 Influence Ranking Evaluation

This section investigates the second research question for influence. In Section 5.3.1, we evaluate the performance of all ranking methods in terms of both flavors of influence. In subsequent sections and for the interest of space, we focus on I-PR as it appears more often in past literature. Another reason is that PageRank-defined influence seems a more reasonable choice, compared to I-CC, to capture the long-term impact of a paper in its discipline. This is because PageRank considers the influence not only of works directly citing the paper of interest, but also works indirectly citing it, through the citation of its citing papers. Section 5.3.2 investigates the effect of the future/present ratio, and Section 5.3.3 examines ranking effectiveness on different top-ranked sets.

5.3.1 Overview of Effectiveness

In this experiment, we fix the future/present ratio to its default value ($\eta = 1.6$) and calculate all evaluation metrics (ρ, τ, top-50 precision, nDCG@50) quantifying the effectiveness of all ranking methods against the ground truth for ranking by influence, either based on citation count (I-CC) or PageRank (I-PR). Tables 3–6 present the results per dataset. We observe that, with regards to I-CC, RAM is the best ranking method outperforming others on most datasets and for most metrics. With regards to I-PR, methods PR and CR appear to be the winners, with the former achieving the highest overall correlations (ρ, τ), and the latter performing best in terms of top-50 accuracy (precision, nDCG).

Let us first study the ranking by I-CC. We expect citation count-based methods to perform well in terms of overall correlation, which is indeed the case with RAM and WC. Focusing on the top-50 results though, we observe that RAM is better than WC at distinguishing the most influential papers. As both methods employ time-aware weights in the adjacency matrix, we may conclude that citation age (RAM) is more effective than citation gap (WC). Some further interesting observations can be made for PageRank-based methods. Although CR and FR fail to capture the overall correlation, they perform remarkably well in distinguishing the top-50 papers. We conjecture this is not because of employing PageRank (other such methods are not doing well) but rather in their use of time-aware landing probabilities. Surprisingly, combining time-aware landing probabilities and time-awareness in the adjacency matrix does not appear to help, as shown in the case of NR. Moreover, despite considering citation chains (and not only direct citations), ECM performs equally well to RAM. This is partly because they both use the same type of time-awareness, and partly because ECM heavily attenuates the importance of long citation chains. This way ECM is mainly determined by direct citations and hence it largely approximates RAM. We observe this close relationship between RAM and ECM across all datasets, impact aspects, and evaluation metrics.

We now turn to I-PR, where we expect PageRank-based methods to be at an advantage. This is the case, with PR and NPR being the best methods, with CR and FR achieving remarkable performance for top-k precision and nDCG. Plain PR achieves the best effectiveness in terms of ρ, τ since the general structure of the graph, $\boldsymbol{A}(t_c)$, on which PR runs is largely similar to the graph, $\boldsymbol{A}(t_c + T)$, on which the ground truth PageRank is computed. While PR captures the ranking of all papers well, on average, it is not the best method when we focus on the top-50 results. CR best distinguishes highly ranked papers, thanks to its debiasing mechanism that promotes recently published papers over older ones whose importance is overestimated by PR. FR achieves good performance in these scenarios for the same reason.

14. A level of correlation is expected since, for instance, many influential papers remain relatively popular etc.

TABLE 3: hep-th: metrics for I-CC, I-PR; $\eta = 1.6$, $k = 50$.

hep-th	I-CC				I-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.734	0.571	0.480	0.645	**0.892**	**0.768**	0.780	0.909
NPR	0.675	0.513	0.260	0.482	0.874	0.726	0.480	0.793
CR	0.752	0.571	0.720	0.880	0.822	0.652	**0.880**	**0.967**
FR	0.512	0.380	**0.740**	0.865	0.419	0.300	0.780	0.958
ECM	0.830	0.679	0.400	0.795	0.684	0.509	0.320	0.665
RAM	**0.836**	**0.689**	0.700	**0.946**	0.698	0.524	0.480	0.802
NR	0.307	0.216	0.200	0.470	0.338	0.242	0.300	0.622

TABLE 4: APS: metrics for I-CC, I-PR; $\eta = 1.6$, $k = 50$.

APS	I-CC				I-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	**0.760**	**0.603**	0.300	0.611	**0.897**	**0.781**	0.740	0.955
NPR	0.734	0.576	0.340	0.626	0.888	0.752	0.760	0.952
CR	0.573	0.437	0.600	0.827	0.486	0.357	**0.780**	**0.958**
FR	0.486	0.361	**0.640**	0.840	0.377	0.269	0.620	0.847
ECM	0.692	0.534	0.580	**0.846**	0.577	0.419	0.340	0.664
RAM	0.692	0.534	0.600	0.837	0.576	0.419	0.340	0.663
NR	0.169	0.120	0.200	0.290	0.030	0.022	0.220	0.379

TABLE 5: PMC: metrics for I-CC, I-PR; $\eta = 1.6$, $k = 50$.

PMC	I-CC				I-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.726	0.591	0.360	0.652	**0.818**	**0.694**	0.840	0.969
NPR	0.708	0.570	0.360	0.649	0.814	0.682	0.780	0.952
CR	0.563	0.426	0.580	0.842	0.603	0.457	**0.900**	**0.990**
FR	0.261	0.196	0.580	0.803	0.219	0.161	0.820	0.977
ECM	**0.787**	0.677	0.800	0.967	0.751	0.594	0.440	0.797
RAM	**0.787**	**0.679**	**0.820**	**0.969**	0.751	0.596	0.420	0.794
NR	0.183	0.134	0.360	0.555	0.226	0.160	0.580	0.812
YR	0.614	0.469	0.400	0.693	0.618	0.467	0.760	0.938
WSDM	0.567	0.432	0.160	0.478	0.465	0.326	0.140	0.437
WC	0.772	0.649	0.660	0.895	0.737	0.574	0.380	0.715

TABLE 6: DBLP: metrics for I-CC, I-PR; $\eta = 1.6$, $k = 50$.

DBLP	I-CC				I-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.811	0.673	0.480	0.717	**0.884**	**0.778**	0.820	0.981
NPR	0.797	0.655	0.440	0.726	0.880	0.763	0.740	0.965
CR	0.549	0.413	0.740	0.938	0.537	0.402	**0.860**	**0.988**
FR	0.389	0.294	0.780	0.947	0.349	0.257	0.720	0.950
ECM	**0.845**	0.726	**0.820**	**0.966**	0.812	0.656	0.520	0.800
RAM	**0.845**	**0.727**	**0.820**	**0.966**	0.812	0.656	0.520	0.800
NR	0.101	0.074	0.400	0.710	0.050	0.037	0.440	0.714
YR	0.627	0.490	0.620	0.836	0.682	0.553	0.800	0.970
WSDM	0.616	0.465	0.580	0.698	0.593	0.437	0.440	0.688
WC	0.839	0.714	0.480	0.630	0.833	0.682	0.320	0.499

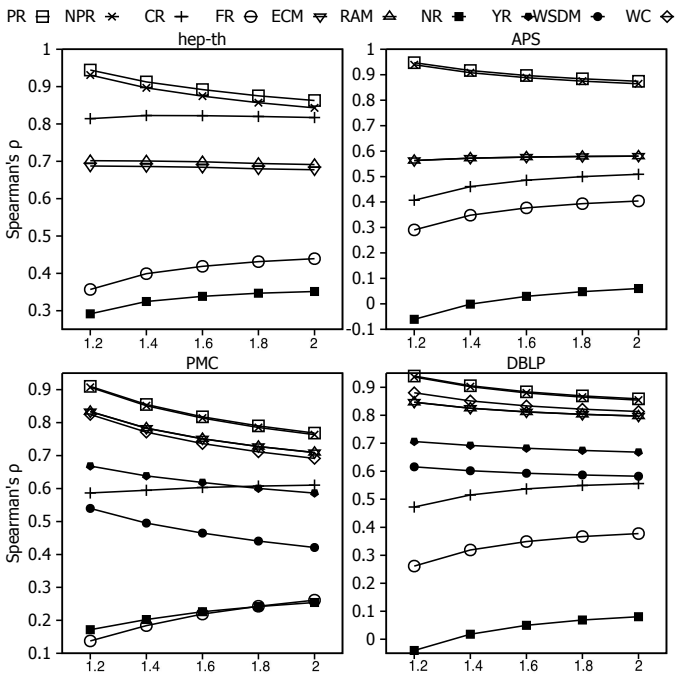

Fig. 1: Correlation of each method's ranking to that of I-PR, varying η.

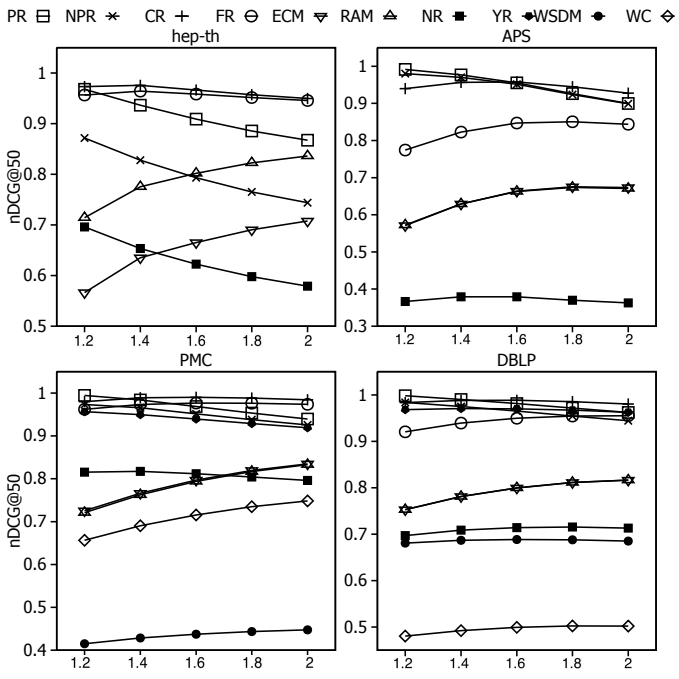

Fig. 2: nDCG@50 of each method's ranking with respect to I-PR, varying η.

5.3.2 Varying the Future/Present Ratio

In this experiment, we vary the future/present ratio η and measure effectiveness of all methods against the ground truth ranking compiled in terms of I-PR. Figure 1 presents Spearman's ρ for each method and dataset; Kendall's τ results are similar and omitted. As also previously observed, PageRank-based methods, and particularly PR and NPR, perform the best in terms of overall correlation; recall that NPR is a basic PR variant that does not introduce time-awareness or use external information. The smaller the ratio η is, the larger the correlation is for these methods. This is because the difference between the present network and the future network, on which the I-PR ranking is derived, is smaller. Interestingly, we observe a group of methods, CR, FR, and NR, that appear to benefit from an increase in η. These PageRank-based methods employ time-aware landing probabilities, and perform rather poorly (except CR) in terms of overall correlation. What happens is the following. As η increases, the ground truth ranking begins to diverge from the simple PageRank-based ranking on the present network, and the primary cause is that recent papers get a chance to gain citations and improve their relative influence with respect to older papers. Time-aware landing probabilities directly account for this phenomenon by promoting *all* recent papers. However, not all recent papers are equally influential, and this explains the relatively small correlations. Increasing the value of η means that the few recent papers that are actually influential will gain visibility in the ranking, and thus the effectiveness of methods that explicitly promote them (along with other recent papers) will increase.

Figure 2 investigates the quality of the top-50 papers returned by each method, and computes their nDCG as the future/present ratio varies; similar findings hold for top-50 precision. The plain PR method is among the best methods, but it loses its crown as η increases. Across the tested η values, CR appears to be the overall best method, often followed by FR. As previously noted, CR, FR and YR employ time-aware landing probabilities to explicitly promote recent papers. When looking at the top ranks, a paper will appear there if it was already influential (in the present network) or if it was moderately influential but attracted many recent citations. CR and FR will identify both types of papers, while PR can only identify the first type. Another interesting observation is that time-aware weighing of the transition/adjacency matrix, used e.g., by RAM, ECM, WC, NR, is not good for discerning PageRank-defined influential papers. Their effectiveness, however, quickly increases with η, indicating a time-aware mechanism is important, but appears to plateau, suggesting that time-aware landing probabilities is a better mechanism.

5.3.3 Varying the Number of Results

In the last experiment for influence-based ranking, we measure the nDCG of each method at various ranks $k \in \{5, 10, 50, 100, 500\}$, as we fix $\eta = 1.6$. Figure 3 presents the results. Overall, we discern that the strong methods, PR, NPR, CR, FR, YR, are robust with respect to k. The effectiveness of the other methods varies greatly with k.

5.4 Popularity Ranking Evaluation

In this section, we investigate the second research question for popularity. In Section 5.4.1, we investigate the effectiveness of all methods with respect to both flavors of popularity. Then, in the following sections, we consider only P-CC, as it is more widely used. Another reason is that for measuring popularity, one might be more interested in measuring the direct impact, as captured by citation count,

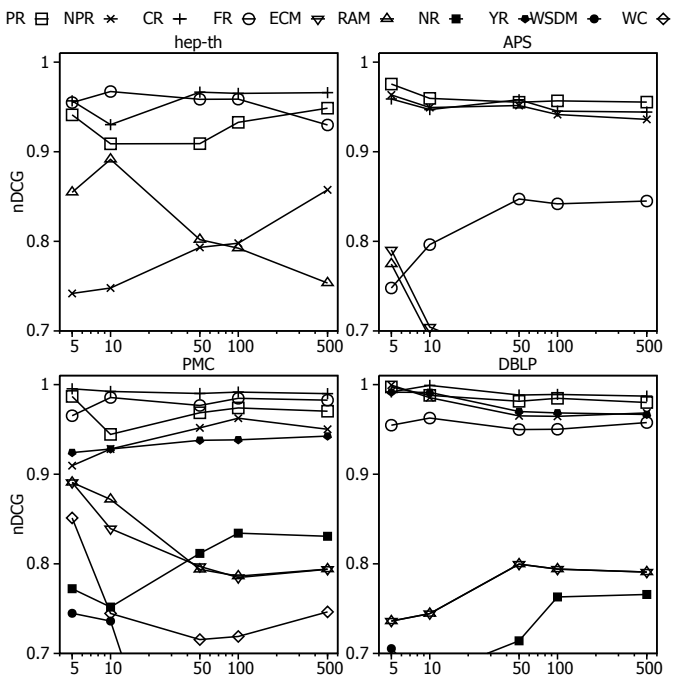

Fig. 3: nDCG of each method's ranking with respect to I-PR, computed at various ranks k; $\eta = 1.6$.

rather than indirect via citation chains. Section 5.4.2 investigates the effect of the future/present ratio, and Section 5.4.3 looks at different ranks.

5.4.1 Overview of Effectiveness

In the first experiment, we fix the future/present ratio to its default value ($\eta = 1.6$) and calculate all evaluation metrics (ρ, τ, top-50 precision, nDCG@50) quantifying the effectiveness of all ranking methods against the ground truth ranking by popularity, either based on citation count (P-CC) or PageRank (P-PR). Tables 7–10 present the results per dataset. In general, we observe that, for P-CC, RAM and ECM perform best for all evaluation metrics, on most datasets. For P-PR, we find that CR achieves the best overall correlation, while RAM, ECM, and FR return a better set of top-50 papers.

Looking into P-CC in more detail, the strength of RAM and ECM is because they mainly rank papers based on their recently received citations, which is a good indication of the citations they will receive in the near-term. An exception appears in the APS dataset, where CR achieves better correlation than RAM and ECM. This phenomenon is due to the nature of the dataset, in that the ranking by P-CC is strongly correlated to that by P-PR (see Table 2), for which, as we discuss next, CR achieves the highest correlation. Overall, we note that time-awareness is important for algorithms to identify popular papers by citations. In contrast to their effectiveness in the case of citation count-defined influence, PageRank methods that employ time-aware landing probabilities (particularly, CR and FR) result in good rankings for citation count-defined popularity.

With respect to P-PR, the best overall correlation is by CR. This is because the focused researcher behavior assumed by CR matches the process by which papers are ranked in the P-PR ground truth. Recall that CR assumes that researchers prefer to read recent papers, while P-PR puts emphasis on papers in the near future. Because recent and near future papers are published nearby in time, their citation trends tend to be alike and, thus, the two processes generate similar rankings. We note, however, that the correlations are significantly weaker than those in the case of PageRank-defined influence. Regarding the top-50 accuracy of methods, we observe that the citation age-weighted adjacency matrix methods, RAM and ECM, perform best, closely followed by the time-aware landing probability methods FR and CR. Similar to time-aware landing probabilities, citation age-weighting biases the simulated researcher towards reading recent papers.

Overall, we observe that removing the bias in favor of old papers is critical for assessing the popularity-based impact of papers. We clearly see that non-time aware methods (PR, NPR, WSDM) do not perform well. However, we also observe that not all types of time-awareness help. For instance, NR overcompensates the bias, as it adjusts both the adjacency matrix and the landing probabilities, simulating

TABLE 7: hep-th: metrics for P-CC, P-PR; $\eta = 1.6$, $k = 50$.

hep-th	P-CC				P-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.301	0.217	0.300	0.332	0.243	0.166	0.240	0.353
NPR	0.249	0.179	0.200	0.240	0.227	0.161	0.180	0.410
CR	0.561	0.416	0.500	0.638	**0.466**	**0.323**	0.400	0.566
FR	0.548	0.407	0.540	0.659	0.453	0.313	**0.480**	0.620
ECM	0.578	0.437	0.400	0.776	0.319	0.219	0.360	0.531
RAM	**0.601**	**0.460**	**0.580**	**0.855**	0.360	0.251	0.440	**0.640**
NR	0.311	0.223	0.200	0.370	0.339	0.231	0.220	0.494

TABLE 8: APS: metrics for P-CC, P-PR; $\eta = 1.6$, $k = 50$.

APS	P-CC				P-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.159	0.113	0.160	0.347	0.127	0.085	0.120	0.347
NPR	0.153	0.109	0.220	0.359	0.134	0.090	0.180	0.378
CR	**0.570**	**0.423**	0.500	0.627	**0.529**	**0.371**	0.420	0.642
FR	0.554	0.412	0.540	0.658	0.518	0.361	**0.440**	0.653
ECM	0.500	0.377	0.540	**0.716**	0.399	0.280	0.420	**0.672**
RAM	0.509	0.385	**0.580**	0.705	0.412	0.289	**0.440**	0.667
NR	0.356	0.255	0.160	0.199	0.354	0.240	0.220	0.308

TABLE 9: PMC: metrics for P-CC, P-PR; $\eta = 1.6$, $k = 50$.

PMC	P-CC				P-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.332	0.260	0.220	0.421	0.198	0.141	0.260	0.444
NPR	0.319	0.250	0.220	0.427	0.200	0.142	0.280	0.484
CR	0.412	0.316	0.360	0.648	**0.272**	**0.189**	0.440	0.717
FR	0.357	0.268	0.400	0.658	0.255	0.174	0.520	**0.818**
ECM	**0.435**	**0.350**	0.660	0.896	0.224	0.161	0.540	0.772
RAM	0.434	**0.350**	**0.680**	**0.902**	0.226	0.163	**0.560**	0.786
NR	0.255	0.193	0.300	0.482	0.245	0.170	0.520	0.767
YR	0.335	0.249	0.220	0.461	0.125	0.084	0.240	0.448
WSDM	0.385	0.291	0.140	0.382	0.041	0.027	0.140	0.317
WC	0.399	0.316	0.500	0.745	0.159	0.113	0.400	0.565

TABLE 10: DBLP: metrics for P-CC, P-PR; $\eta = 1.6$, $k = 50$.

DBLP	P-CC				P-PR			
	ρ	τ	prec	nDCG	ρ	τ	prec	nDCG
PR	0.347	0.257	0.160	0.384	0.279	0.194	0.200	0.449
NPR	0.336	0.248	0.160	0.382	0.282	0.196	0.220	0.458
CR	0.533	0.397	0.440	0.717	**0.496**	**0.348**	0.500	0.765
FR	0.454	0.337	0.480	0.740	0.446	0.311	0.520	0.788
ECM	**0.552**	**0.428**	**0.700**	**0.886**	0.433	0.310	**0.620**	**0.855**
RAM	0.550	0.427	0.680	0.876	0.432	0.310	**0.620**	0.846
NR	0.262	0.190	0.300	0.561	0.310	0.212	0.420	0.684
YR	0.352	0.256	0.320	0.537	0.287	0.196	0.340	0.595
WSDM	0.314	0.227	0.320	0.442	0.145	0.097	0.300	0.464
WC	0.421	0.315	0.320	0.447	0.282	0.196	0.240	0.431

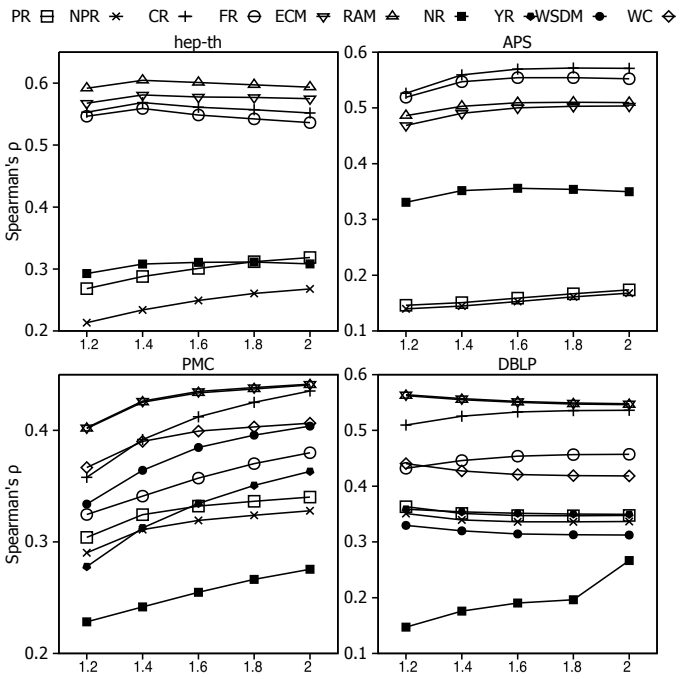

Fig. 4: Correlation of each method's ranking to that of P-CC, varying η.

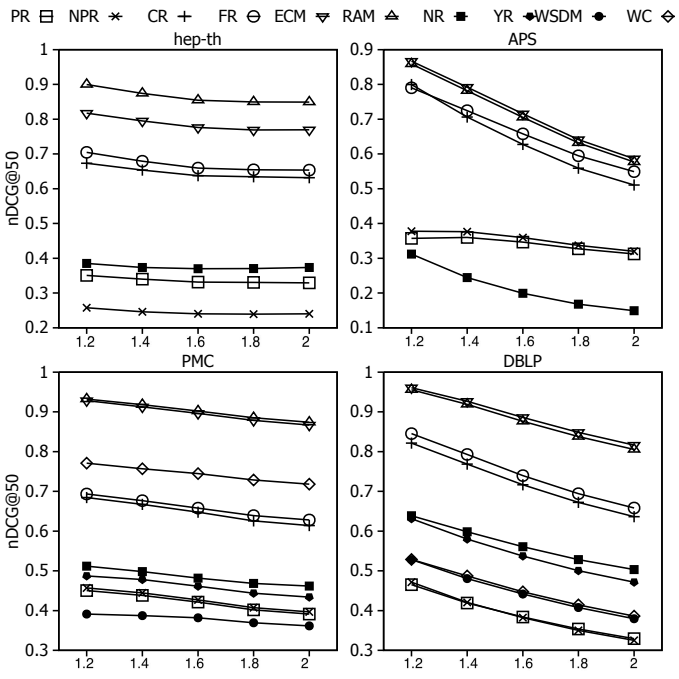

Fig. 5: nDCG@50 of each method's ranking with respect to P-CC, varying η.

thus a researcher that starts at recent papers *and* prefers following references to recent papers. On the other hand, WC chooses to promote papers that were cited quickly (small citation gap), regardless of whether this happened recently or not, failing thus to capture current network dynamics.

5.4.2 Varying the Future/Present Ratio

We next vary the future/present ratio η and measure effectiveness of all methods against the ground truth ranking by P-CC. Figure 4 presents Spearman's ρ for each method and dataset; Kendall's τ results are similar and omitted. The general observation of the previous section applies here. RAM, ECM, and CR are the strongest methods on all datasets independently of η.

As η increases, we observe that the effectiveness of the methods increases at first, but then plateaus and subsequently decreases. There are two forces at play here. To understand the first, recall that methods extrapolate from the citation trends in the recent history to compile the rank of papers in the future. It is thus natural to expect that as η increases, at some point, the accuracy of this extrapolation will begin to deteriorate. This is why for large η values we observe plateaus and/or decreases in correlation.

For the second force, note that the distribution of citations follows a "power law", or a closely related distribution [65], [66]. This means that the vast majority of papers receives few citations, forming the so-called long "tail" of the distribution, while the top-cited papers comprise its much shorter "head". Small η values correspond to a short time period, meaning that there is less chance for the papers in the tail to gather enough citations to differentiate among themselves — any reported differences between them may be coincidental. Hence, their ranking in the ground truth is based on a small sample of citations that is not representative of their relative importance. All methods are bound to have trouble distinguishing between the papers in the tail, and since they are significantly more numerous, we expect low correlations for small η values. Increasing η alleviates the previous issue and the performance of all methods improves.

Figure 5 presents the results for nDCG@50, where the relative ordering among the methods is preserved as η varies; similar findings hold for top-50 precision. The most interesting observation is that all methods have trouble identifying the popular (in terms of citation count) papers as we look further in the future. The reason is that in this experiment we focus on top papers that are being heavily cited, meaning that the aforementioned second force does not apply here. As a result, the underlying citation distribution of these top papers is evident even for small η values and, thus, varying η reveals only the divergence of their citation trends as $t_c + T$ departs from t_c.

5.4.3 Varying the Number of Results

In the last experiment, we measure the nDCG of each method at various ranks k, while we fix the future/present ratio to 1.6. Figure 6 presents the results. We observe that the strong methods, RAM and ECM, consistently identify papers with high popularity-based impact at all tested ranks. It is worth noting, that while methods cannot correctly estimate a ranking that is overall similar to that of P-CC, they do have the ability to push the most highly cited papers, based on P-CC, to their top ranks.

5.5 Convergence Rate

In this section, we compare the convergence rate of the iterative methods, based on their parameter settings in the

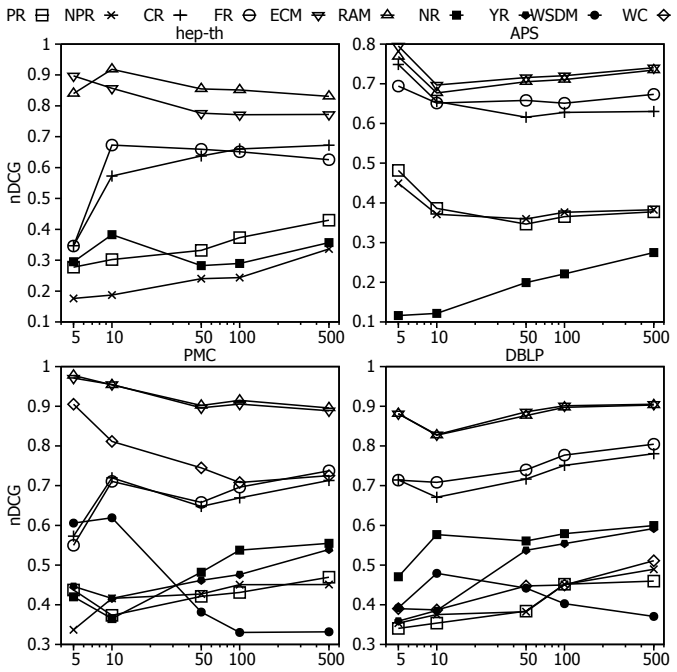

Fig. 6: nDCG of each method's ranking with respect to P-CC, computed at various ranks k; $\eta = 1.6$.

I-PR and P-CC scenarios as presented in Sections 5.3.1 and 5.4.1. In particular, we chose the best parameter settings in terms of effectiveness based on Spearman's ρ. Figures 7 and 8 present the convergence error of each method, per dataset, as a function of running time, for I-PR and P-CC, respectively. In each scenario, we present the methods ordered based on their effectiveness, i.e., their achieved Spearman's ρ.

In all scenarios ECM is the quickest method to converge. ECM is based on the Katz centrality (see Section 3.2), which weights a node on the graph based on all paths that pass through it [51]. ECM's quick convergence is due to its parameter setting, which heavily attenuates the weights of paths that are longer than direct citations. The convergence of PR and its variants (NPR, NR, CR,and YR) depends on the random jump probability α (see Equations 1-2), with values closer to 1 increasing the number of iterations required for convergence (and, hence, running time) [16]. We observe that of these methods PR, CR, NR, and YR converge in comparable time. This is because they all use similar values of $\alpha \sim 0.5$. An exception occurs for YR on the I-PR scenario, where YR converges much slower. This is because its best setting on DBLP uses $\alpha = 0.85$, compared to $\alpha = 0.5$ on PMC. An interesting observation is that while NPR is also a simple PR variant, with a similar α value, the non-linear nature of its calculations causes it to converge significantly slower. In particular, NPR requires about $15 - 25$ more iterations compared to simple PR, hence its increased running time. Finally, FR converges slightly slower compared to simple PR. It is noteworthy that this is not related to the number of iterations required, but due to its heavier calculations per iteration, since it uses multiple networks.

To provide a direct comparison of running time trade-offs between methods on each dataset, we present in Figures 9 and 10 the running time required to achieve a convergence error of less than 10^{-12} in the I-PR and P-CC scenarios, respectively. When running time is an issue, PR is the preferred method for I-PR, as it is quite fast (only ECM is faster) and has the highest effectiveness. For P-CC, ECM is clearly the preferred method, being both the fastest and the most effective.

TABLE 11: State-of-the-art method in terms of overall correlation/top-k ranking.

	Influence	**Popularity**
Citation Count	RAM/RAM	ECM/RAM
PageRank	PR/CR	CR/RAM

5.6 Discussion

Our evaluation was based on four real datasets and focused on answering three questions, (1) how distinct are the notions of influence and popularity, (2) which method is the state-of-the-art for each impact aspect, and (3) which (iterative) method runs the fastest. In what follows, we summarize our findings about these questions, and then draw some general conclusions about the current state of research.

Regarding the first question, we observe that popularity and influence carry rather distinct semantics, but are to some extent correlated. The correlations are stronger between the popularity (P-CC and P-PR) and between the influence flavors (I-CC and I-PR). Moreover, as we look further into the future (large η values), the similarity between popularity- and influence-based rankings increases. This observation raises the issue of appropriately selecting a suitable time horizon T (or η value) so that the ground truth actually captures popularity. This horizon should be on the one hand short to avoid conflation of short- and long-term impact, and on the other hand long enough to capture the typical duration of the research cycle in the scientific discipline. We also note that the two flavors of influence are strongly correlated for large values of η, suggesting that it makes sense to focus on and optimize for one of them, e.g., I-PR, as past work has done.

Regarding the second question about which method performs best we find no single winner that can identify papers with high impact both in terms of influence and popularity, as expected. Instead, in many cases we see that the choice depends on how impact is defined, but also on what the objective is, i.e., to derive an overall accurate ranking, or to identify the few papers with the highest impact. Table 11 summarizes our findings about which method performed best per setting.

For influence, we draw the following conclusions about what type of approaches work. In general, time-aware weighting of the adjacency matrix is a robust mechanism for identifying influential papers. Time-awareness in landing probabilities has a bipolar effect, reducing the overall correlation dramatically, but helping identify the top most influential papers, particularly in the case of PageRank-defined influence (I-PR), which is the most commonly used ground truth for influence in the literature. Unmodified PageRank

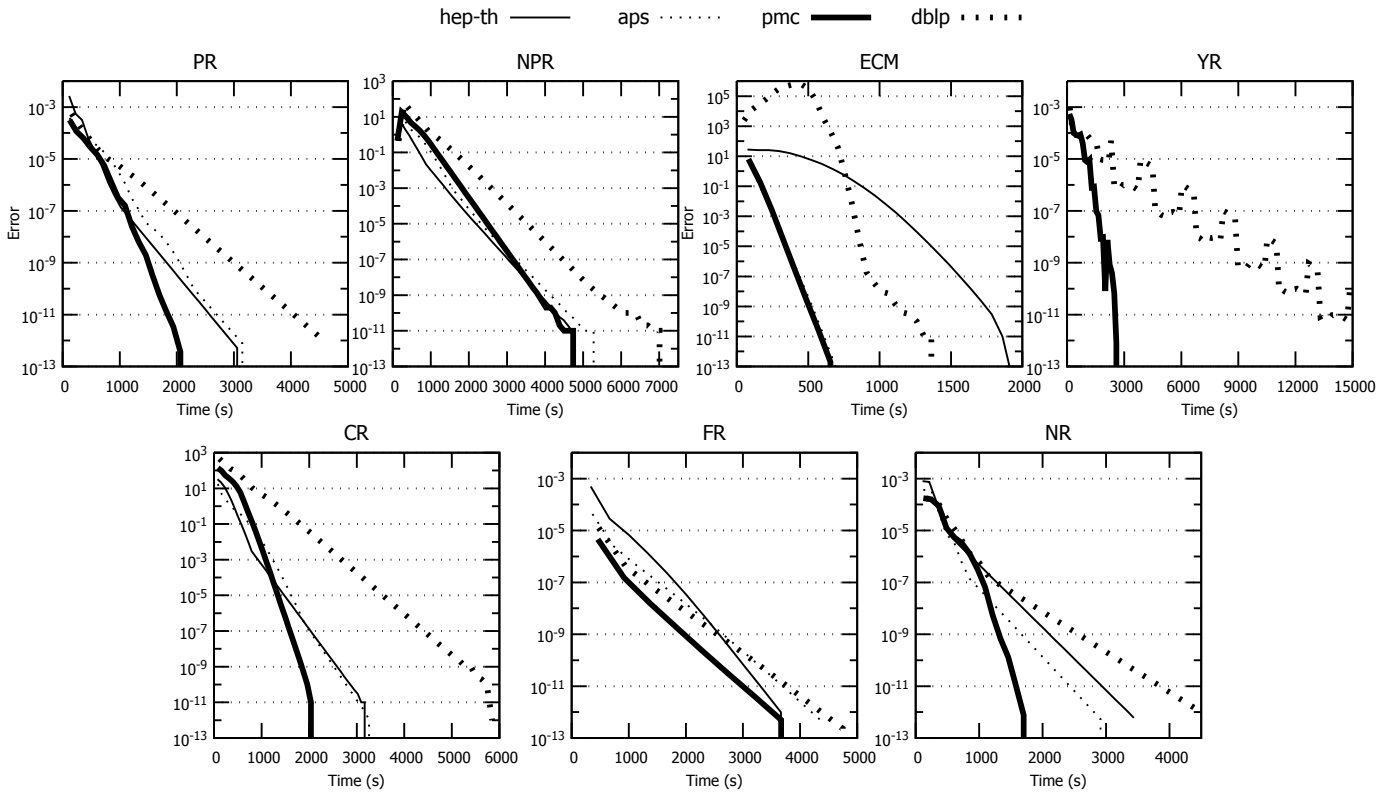

Fig. 7: Convergence Rate per method and dataset based on the optimal I-PR parameter setting.

does a good job in capturing the overall correlation with respect to PageRank-defined influence. Metadata, processes on multiple networks, and ensembles do not appear to help. It should also be noted that, naturally, methods based on one centrality metric (citation count or PageRank) perform better when influence is measured in terms of the corresponding centrality.

Regarding popularity, our conclusions are the following. Time awareness is a vital factor for discerning popularity, as it effectively compensates for the bias in favor of old papers. Overall in the popularity scenario, we again find that metadata, processes on multiple networks, and ensembles do not appear to help. We also note that ranking by popularity appears to be a harder problem than ranking by influence. The performance of all methods is comparatively much lower, especially for larger time horizons. For instance we note that the maximum top-50 precision achieved for popularity is 0.7, whereas it is 0.9 for influence. This observation justifies why the majority of past work has focused on popularity-defined impact. The current performance of the state-of-the-art, however, leaves open the space for further improvements in this direction.

Regarding the third research question, among CR and PR, which are the most effective in the influence scenario, PR could be the preferable choice in scenarios with stricter execution time constraints, since it converges faster in all cases. As regards popularity, where CR and RAM/ECM are the most effective methods, RAM/ECM could be the most viable solution, when taking time constraints into consideration.

6 Future Research Directions

In the following, we discuss how current research could be expanded in the future, based on suggestions from the literature, as well as our own insights.

Popularity Based Ranking improvements. Our evaluation results in Section 5.4 have shown room for improvement with regards to popularity-based ranking, in contrast to influence-based ranking that does not appear to be as hard a problem. Future ranking methods need to address this shortcoming. Improved ranking by popularity could be achieved, for example, by further tackling the "cold start" problem. This problem refers to the fact that new papers have zero, or low citation counts at the time of ranking, although they do get cited shortly after their publication. As we have discussed (see Section 3), some existing ranking methods address this issue by using age-based weights. However, these weights are uniformly determined based on paper, or citation age. As discussed in [25], [46] future research needs to differentiate between cold-start papers since not all recent papers attract citations, while those that do are not all cited at the same rate. Identifying the characteristics which affect citation rate, and which divide papers into those attracting vs. those not attracting citations, and exploiting these characteristics to rank papers is an open issue. A promising step in this direction could be the adoption of network evolution models, proposed by network science theory [67], which has shown that various real networks of different domains exhibit particular common behaviours [68], [69], [70], [71]. Such ideas could help differentiate between cold start papers, and accordingly modify ranking

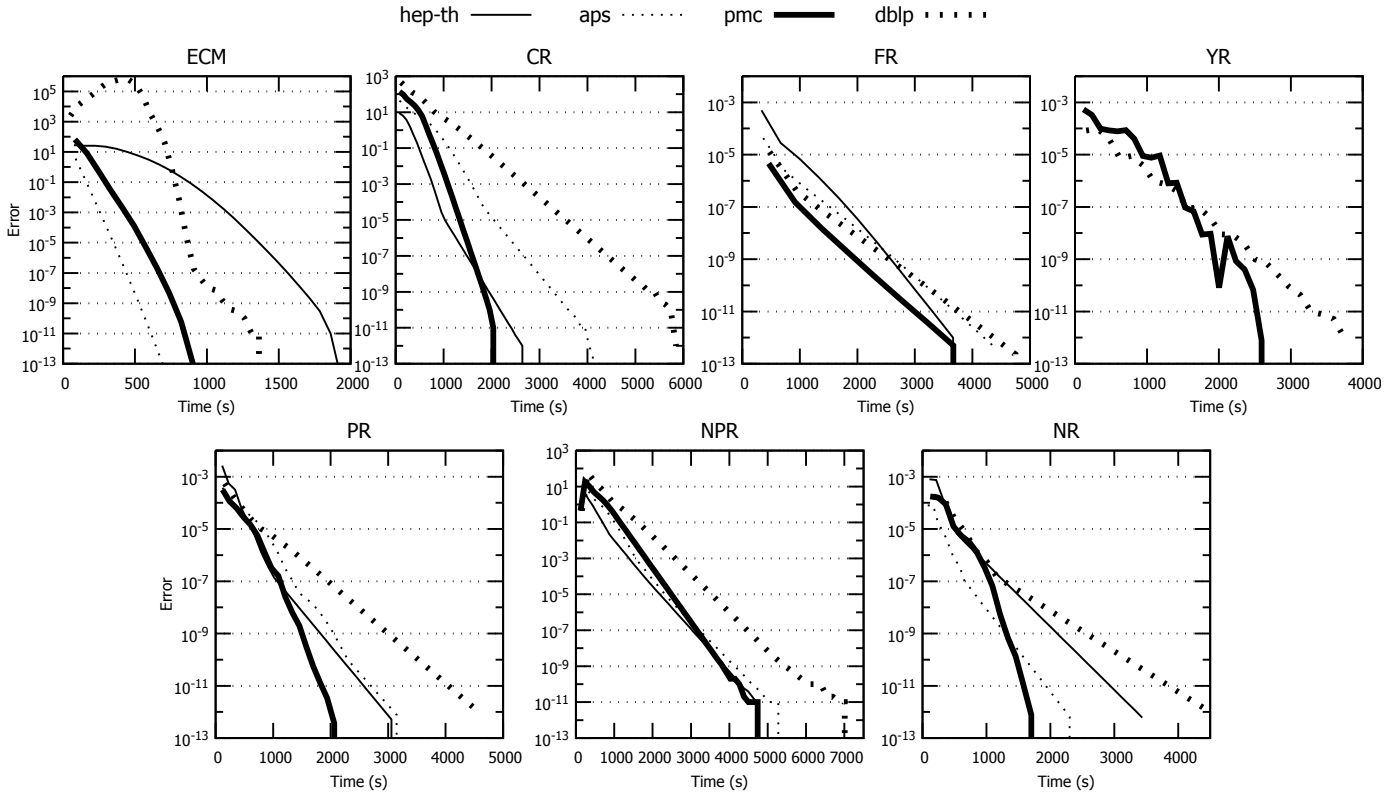

Fig. 8: Convergence Rate per method and dataset based on the optimal parameter setting for P-CC.

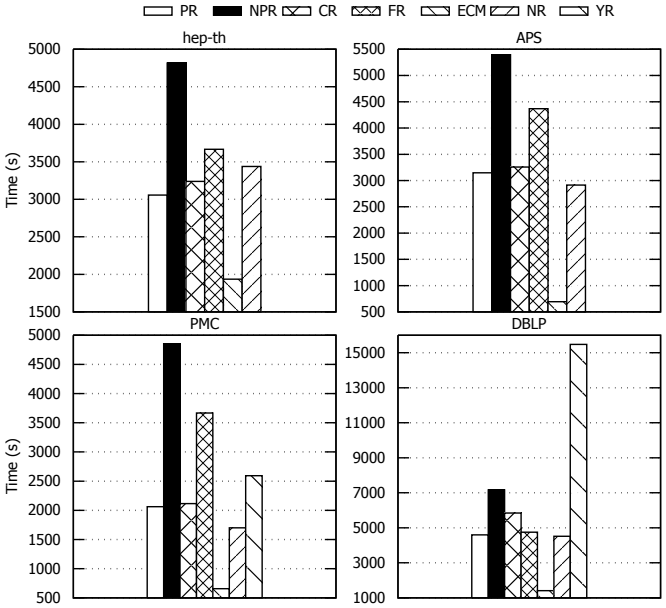

Fig. 9: Running times per method based on the optimal parameter setting for I-PR.

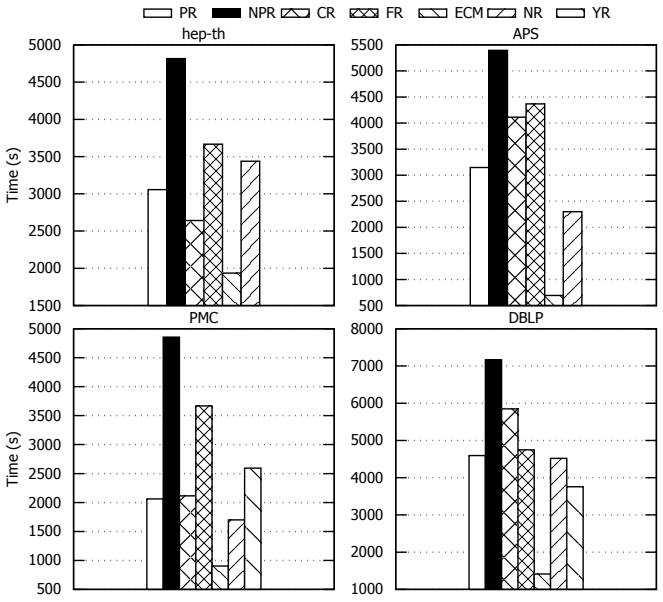

Fig. 10: Running times per method based on the optimal parameter setting for P-CC.

models (e.g., modifying PageRank's Random Surfer Model), to achieve improved popularity-based ranking.

Metadata exploitation. Existing work has considered using author and venue information in their ranking models, as well as time-related information. As we have shown in our evaluation, however, such methods do not seem to gain a significant advantage in ranking based on popularity or influence. This, of course, does not imply that using article metadata is meaningless. One direction for better exploiting article metadata could be via Machine Learning (ML) techniques. For example, in [31], [45], [46] the authors propose that future work could combine network analysis with ML techniques, e.g., to learn paper weights that

can be incorporated into a ranking formula. A particularly promising approach in this direction would be to exploit citation count prediction techniques [72], [73], [74]. These methods use learning techniques over article metadata to predict future paper citations and, hence, their output could be a useful component of future paper ranking methods.

Moreover, new types of data that have previously not been readily available, could be used. An interesting direction for future work (e.g., as suggested in [9]) would be the use of Altmetrics [75], [76], [77]. Altmetrics are data derived from web usage statistics (e.g., paper downloads, views, etc). These data have not been exploited for impact-based ranking so far, and could be incorporated in the mechanisms of paper ranking methods (e.g., as parts of paper, or citation weights).

Benchmarks and metrics. An open research direction concerns the standardisation of methodologies to evaluate paper ranking methods, i.e., formulating commonly accepted measures for their evaluation [8], [9], [51]. In our work, we have presented formal definitions for popularity and influence, two impact aspects that are often conflated in current literature, and presented a framework to evaluate ranking methods based on their ranking capability with regards to these impact aspects. However, it would also be interesting to rank scientific papers in terms of other interesting properties (such as their content's novelty), apart from their impact. Methodically examining and ranking articles based on such properties is, of course, inextricably linked to formally describing these properties and additionally devising frameworks to evaluate ranking methods based on them.

Dataset compilation. A major deficiency of previous work is the lack of evaluations conducted on multiple datasets. Most works so far have only performed evaluations on few, if not only a single dataset. In this work, we have performed experiments on four datasets of three different scientific domains, tackling this deficiency. However, evaluations should be reproducible on additional datasets of different sizes. Additionally, the move towards open science [78] is establishing the reproducibility of scientific research as a key tenet of the scientific process. To this end, the distribution of datasets and methods becomes an important part of the research cycle. Hence, additional, larger (and interdisciplinary) datasets need to be compiled and made available to the scientific community as part of future research.

Addressing malicious behaviour. The world of research and academia is not flawless: the use of quantitative measures in the assessment of academics and researchers has also brought about malpractices by some researchers aiming to promote their own work. In particular self citation abuse is a widely recognised problem, which has even led to the proposal of metrics such as the S-index [79], a self citation index inspired by the widely used h-index. Another misuse is that of mutual citations between scientists, based on previous collaboration, or acquaintance rather than academic merit. These issues of citation abuse have been brought up by previous works on paper ranking methods [8], [24]. While some existing works are motivated by the need to address self-citation malpractices (e.g., [38]), or address mutual citations (e.g., [41]), open questions still exist. For example, how robust are paper ranking methods against these practices, and how can valid self-citations, or mutual citations (i.e., actually citing relevant work by the same author, or her collaborators) be differentiated from those that are made abusively.

7 Conclusion

In this paper, we formalized the problem of ranking articles based on two distinct impact aspects, influence and popularity. We presented a broad overview and categorization of paper ranking methods proposed in the literature, focusing on what mechanisms have been proposed. We also summarized the evaluation approaches that have been used. Our evaluation was based on four real datasets and focused on answering three research questions: (1) how distinct are the notions of influence and popularity, (2) which method is the state-of-the-art for each impact aspect, and (3) how quickly do iterative methods converge and run.

Our results have shown that for influence PR and CR yield the best ranking results, while for popularity RAM and ECM are the winners with regards to ranking. ECM has been shown to be the quickest method to run in all scenarios. Our evaluation has shown that, overall, ranking by influence is adequately addressed by existing methods. In contrast, we identified a performance gap in the case of ranking by popularity, which showcases the need to develop novel ranking methods to close it. As an epilogue, we pinpointed possible directions for future work in the field of impact-based ranking of papers.

Acknowledgements

We acknowledge support of this work by the project "Moving from Big Data Management to Data Science" (MIS 5002437/3) which is implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

References

[1] P. Larsen and M. Von Ins, "The rate of growth in scientific publication and the decline in coverage provided by science citation index," *Scientometrics*, vol. 84, no. 3, pp. 575–603, 2010.

[2] L. Bornmann and R. Mutz, "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, 2015.

[3] *UNESCO Science Report: towards 2030*. UNESCO Publishing, 2015.

[4] D. Fanelli, "Do pressures to publish increase scientists' bias? an empirical support from us states data," *PLOS ONE*, vol. 5, no. 4, pp. 1–7, 04 2010. [Online]. Available: https://doi.org/10.1371/journal.pone.0010271

[5] D. Sarewitz, "The pressure to publish pushes down quality," *Nature*, vol. 533, no. 7602, pp. 147–147, 2016.

[6] J. P. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, no. 8, p. e124, 2005.

[7] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, "A principal component analysis of 39 scientific impact measures," *PloS one*, vol. 4, no. 6, p. e6022, 2009.

[8] M. Dunaiski and W. Visser, "Comparing paper ranking algorithms," in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. ACM, 2012, pp. 21–30.

[9] X. Bai, H. Liu, F. Zhang, Z. Ning, X. Kong, I. Lee, and F. Xia, "An overview on evaluating and predicting scholarly article impact," *Information*, vol. 8, no. 3, p. 73, 2017.

[10] A. Sidiropoulos and Y. Manolopoulos, "Generalized comparison of graph-based ranking algorithms for publications and authors," *Journal of Systems and Software*, vol. 79, no. 12, pp. 1679–1700, 2006.

[11] S. Ma, C. Gong, R. Hu, D. Luo, C. Hu, and J. Huai, "Query independent scholarly article ranking," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE, 2018, pp. 953–964.

[12] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1, pp. 107–117, 1998.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." 1999.

[14] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with google's pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.

[15] N. Ma, J. Guan, and Y. Zhao, "Bringing pagerank to the citation analysis," *Information Processing & Management*, vol. 44, no. 2, pp. 800–810, 2008.

[16] A. N. Langville and C. D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[17] W.-S. Hwang, S.-M. Chae, S.-W. Kim, and G. Woo, "Yet another paper ranking algorithm advocating recent publications," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1117–1118.

[18] P. S. Yu, X. Li, and B. Liu, "On the temporal dimension of search," in *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*. ACM, 2004, pp. 448–449.

[19] V. P. Diodato and P. Gellatly, *Dictionary of Bibliometrics (Haworth Library and Information Science)*. Routledge, 1994. [Online]. Available: https://www.amazon.com/Dictionary-Bibliometrics-Haworth-Library-Information/dp/1560248521?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=1560248521

[20] E. V. Bernstam, J. R. Herskovic, Y. Aphinyanaphongs, C. F. Aliferis, M. G. Sriram, and W. R. Hersh, "Using citation data to improve retrieval from medline," *Journal of the American Medical Informatics Association*, vol. 13, no. 1, pp. 96–105, 2006.

[21] P. Groth and T. Gurney, "Studying scientific discourse on the web using bibliometrics: A chemistry blogging case study," 2010.

[22] D. R. Smith, "A 30-year citation analysis of bibliometric trends at the archives of environmental health, 1975–2004," *Archives of environmental & occupational health*, vol. 64, no. sup1, pp. 43–54, 2009.

[23] D. Wang, C. Song, and A. Barabási, "Quantifying long-term scientific impact," *CoRR*, vol. abs/1306.3293, 2013. [Online]. Available: http://arxiv.org/abs/1306.3293

[24] L. Yao, T. Wei, A. Zeng, Y. Fan, and Z. Di, "Ranking scientific publications: the effect of nonlinearity," *Scientific reports*, vol. 4, 2014.

[25] J. Zhou, A. Zeng, Y. Fan, and Z. Di, "Ranking scientific publications with similarity-preferential mechanism," *Scientometrics*, vol. 106, no. 2, pp. 805–816, 2016.

[26] M. Krapivin and M. Marchese, "Focused page rank in scientific papers ranking," in *International Conference on Asian Digital Libraries*. Springer, 2008, pp. 144–153.

[27] C. Su, Y. Pan, Y. Zhen, Z. Ma, J. Yuan, H. Guo, Z. Yu, C. Ma, and Y. Wu, "Prestigerank: A new evaluation method for papers and journals," *Journal of Informetrics*, vol. 5, no. 1, pp. 1–13, 2011.

[28] E. Yan and Y. Ding, "Weighted citation: An indicator of an article's prestige," *Journal of the American Society for Information Science and Technology*, vol. 61, pp. 1635–1643, August 2010.

[29] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman, "Time-aware ranking in dynamic citation networks," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 373–380.

[30] P. S. Yu, X. Li, and B. Liu, "Adding the temporal dimension to search-a case study in publication search," in *Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on*. IEEE, 2005, pp. 543–549.

[31] H. Chin-Chi, C. Kuan-Hou, F. Ming-Han, W. Yueh-Hua, C. Huan-Yuan, Y. Sz-Han, C. Chun-Wei, T. Ming-Feng, Y. Mi-Yen, and L. Shou-De, "Time-aware weighted pagerank for paper ranking in academic graphs," *WSDM Cup*, 2016.

[32] D. Luo, C. Gong, R. Hu, L. Duan, and S. Ma, "Ensemble enabled weighted pagerank," *arXiv preprint arXiv:1604.05462*, 2016.

[33] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, no. 06, p. P06010, 2007.

[34] H. Sayyadi and L. Getoor, "Futurerank: Ranking scientific articles by predicting their future pagerank." in *SDM*. SIAM, 2009, pp. 533–544.

[35] F. Zhang and S. Wu, "Ranking scientific papers and venues in heterogeneous academic networks by mutual reinforcement," in *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*. ACM, 2018, pp. 127–130.

[36] E. Yan, Y. Ding, and C. R. Sugimoto, "P-rank: An indicator measuring prestige in heterogeneous scholarly networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 467–477, 2011.

[37] Y. Wang, Y. Tong, and M. Zeng, "Ranking scientific articles by exploiting citations, authors, journals, and time information." in *AAAI*, 2013.

[38] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, "Identifying anomalous citations for objective evaluation of scholarly article impact," *PloS one*, vol. 11, no. 9, p. e0162364, 2016.

[39] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma, "Object-level ranking: bringing order to web objects," in *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005, pp. 567–574.

[40] X. Jiang, X. Sun, and H. Zhuge, "Towards an effective and unbiased ranking of scientific literature through mutual reinforcement," in *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2012, pp. 714–723.

[41] Z. Liu, H. Huang, X. Wei, and X. Mao, "Tri-rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce," in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. IEEE, 2014, pp. 493–500.

[42] M. Feng, K. Chan, H. Chen, M. Tsai, M. Yeh, and S. Lin, "An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016," *WSDM Cup*, 2016.

[43] S. Chang, S. Go, Y. Wu, Y. Lee, C. Lai, S. Yu, C. Chen, H. Chen, M. Tsai, M. Yeh, and S. Lin, "An ensemble of ranking strategies for static rank prediction in a large heterogeneous graph," *WSDM Cup*, 2016.

[44] D. Herrmannova and P. Knoth, "Simple yet effective methods for large-scale scholarly publication ranking," *arXiv preprint arXiv:1611.05222*, 2016.

[45] I. Wesley-Smith, C. T. Bergstrom, and J. D. West, "Static ranking of scholarly papers using article-level eigenfactor (alef)," *arXiv preprint arXiv:1606.08534*, 2016.

[46] S. Ribas, A. Ueda, R. L. Santos, B. Ribeiro-Neto, and N. Ziviani, "Simplified relative citation ratio for static paper ranking: Ufmg/latin at wsdm cup 2016," *arXiv preprint arXiv:1603.01336*, 2016.

[47] D. F. Klosik and S. Bornholdt, "The citation wake of publications detects nobel laureates' papers," *PloS one*, vol. 9, no. 12, p. e113184, 2014.

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

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

[50] X. Bai, J. Hou, H. Du, X. Kong, and F. Xia, "Evaluating the impact of articles with geographical distances between institutions," in *Proceedings of the 26th international conference on world wide web companion*. International World Wide Web Conferences Steering Committee, 2017, pp. 1243–1244.

[51] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, "Ranking in evolving complex networks," *Physics Reports*, vol. 689, pp. 1–54, 2017.

[52] T. H. Haveliwala, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 784–796, 2003.

[53] G. Jeh and J. Widom, "Scaling personalized web search," in *WWW*. ACM, 2003, pp. 271–279.

[54] E. Garfield, "The history and meaning of the journal impact factor," *Jama*, vol. 295, no. 1, pp. 90–93, 2006.

[55] C. T. Bergstrom, J. D. West, and M. A. Wiseman, "The eigenfactor™ metrics," *The Journal of Neuroscience*, vol. 28, no. 45, pp. 11 433–11 434, 2008.

[56] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[57] A. D. Wade, K. Wang, Y. Sun, and A. Gulli, "Wsdm cup 2016: Entity ranking challenge," in *Proceedings of the ninth ACM international conference on web search and data mining*. ACM, 2016, pp. 593–594.

[58] B. I. Hutchins, X. Yuan, J. M. Anderson, and G. M. Santangelo, "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, 2016.

[59] G. D. Paparo and M. Martin-Delgado, "Google in a quantum network," *Scientific reports*, vol. 2, p. 444, 2012.

[60] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD'08*, 2008, pp. 990–998.

[61] V. Koukis, C. Venetsanopoulos, and N. Koziris, "˜ okeanos: Building a cloud, cluster by cluster," *IEEE Internet Computing*, no. 3, pp. 67–71, 2013.

[62] C. Spearman, "The proof and measurement of association between two things," *The American journal of psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[63] M. G. Kendall, "Rank correlation methods," 1955.

[64] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Brooks/Cole, 1996.

[65] M. Brzezinski, "Power laws in citation distributions: Evidence from scopus," *Scientometrics*, vol. 103, no. 1, pp. 213–228, 2015.

[66] Y.-H. Eom and S. Fortunato, "Characterizing and modeling citation dynamics," *PloS one*, vol. 6, no. 9, p. e24926, 2011.

[67] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[68] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, p. aaf5239, 2016.

[69] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.

[70] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabási, "Career on the move: Geography, stratification, and scientific impact," *Scientific reports*, vol. 4, p. 4770, 2014.

[71] F. Hadiji, M. Mladenov, C. Bauckhage, and K. Kersting, "Computer science on the move: inferring migration regularities from the web via compressed label propagation," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[72] C. Castillo, D. Donato, and A. Gionis, "Estimating number of citations using author reputation," in *International Symposium on String Processing and Information Retrieval*. Springer, 2007, pp. 107–117.

[73] R. Yan, J. Tang, X. Liu, D. Shan, and X. Li, "Citation count prediction: learning to estimate future citations for literature," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1247–1252.

[74] L. Weihs and O. Etzioni, "Learning to predict citation-based impact measures," in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*. IEEE Press, 2017, pp. 49–58.

[75] J. Priem, D. Taraborelli, P. Groth, and C. Neylon, "Altmetrics: A manifesto," 2010.

[76] H. Piwowar, "Introduction altmetrics: What, why and where?" *Bulletin of the American Society for Information Science and Technology*, vol. 39, no. 4, pp. 8–9, 2013.

[77] M. Fenner, "Altmetrics and other novel measures for scientific impact," in *Opening science*. Springer, 2014, pp. 179–189.

[78] A. Farnham, C. Kurz, M. A. Öztürk, M. Solbiati, O. Myllyntaus, J. Meekes, T. M. Pham, C. Paz, M. Langiewicz, S. Andrews *et al.*, "Early career researchers want open science," *Genome biology*, vol. 18, no. 1, p. 221, 2017.

[79] J. Flatt, A. Blasimme, and E. Vayena, "Improving the measurement of scientific success by reporting a self-citation index," *Publications*, vol. 5, no. 3, p. 20, 2017.

Ilias Kanellos is a Ph.D. student at the Electrical and Computer Engineering department of NTUA, under the supervision of prof. Yannis Vassiliou, and a research assistant at IMSI - "Athena" RC. His research interests include scientific data management, information retrieval from scientific manuscripts, cloud computing, and Paper ranking algorithms.

Thanasis Vergoulis is a Scientific Associate at IMSI, Athena Research Center in Greece. He received his diploma in Computer Engineering and Informatics from the Univ. of Patras and his PhD in Computer Science from NTU of Athens, under the supervision of Prof. Timos Sellis. He has been teaching courses in undergraduate and postgraduate level in academic institutions in Greece and Cyprus. His research interests consist of data management, bioinformatics, cloud computing, and research analytics.

Dimitris Sacharidis is an Assistant Professor at the Institute of Information Systems Engineering of TU Wien. Prior to that, he was a junior researcher at IMSI - "Athena" RC, and a Marie Curie postdoctoral fellow at HKUST. His research interests include spatio-temporal and social data analytics, and recommender systems.

Theodore Dalamagas is a Senior Researcher at IMSI - "Athena" RC. He received his Diploma in Electrical Engineering from NTUA (1996), Greece, his MSc in Advanced Information Systems from Glasgow University, Scotland (1997), and his PhD from NTUA (2004). Since 2007, he works at "Athena" Research Center. His research interests include: Scientific databases, Web of data, Information retrieval, Data semantics, and Query processing.

Yannis Vassiliou is a professor Emeritus at the Division of Computer Science of the Electrical and Computer Engineering school of NTUA. His research interests include: Information Systems, Database Management, Information Systems Applications, Data Warehouses, Information Storage and Retrieval, Information Interfaces and Presentation, Models and Principles, Data Engineering, Intelligent Systems, and Software Engineering