Noname manuscript No.
(will be inserted by the editor)

Ranking Papers by their Short-Term Scientific Impact

Ilias Kanellos[1] · Thanasis Vergoulis[1] · Dimitris Sacharidis[3] · Theodore Dalamagas[1] · Yannis Vassiliou[2]

the date of receipt and acceptance should be inserted later

Abstract The constantly increasing rate at which scientific papers are published makes it difficult for researchers to identify papers that currently impact the research field of their interest. Hence, approaches to effectively identify papers of high impact have attracted great attention in the past. In this work, we present a method that seeks to rank papers based on their estimated short-term impact, as measured by the number of citations received in the near future. Similar to previous work, our method models a researcher as she explores the paper citation network. The key aspect is that we incorporate an attention-based mechanism, akin to a time-restricted version of preferential attachment, to explicitly capture a researcher's preference to read papers which received a lot of attention recently. A detailed experimental evaluation on four real citation datasets across disciplines, shows that our approach is more effective than previous work in ranking papers based on their short-term impact.

Keywords Citation Networks · Paper Ranking · Data Mining

1 Introduction

Quantifying the importance of scientific publications, colloquially called *papers*, is an important research problem with various applications. For example, a student that wants to familiarize herself with a research area, may look for seminal papers in the field. A hiring committee may assess an applicant based on the aggregate impact of his publication record. As the number of papers published grows at an increasing rate [19,4], discerning the important papers, especially among the recent publications, becomes a hard task.

Conventionally, the importance of a paper is bestowed upon itself by its peers, the subsequent papers that continue the line of research and acknowledge its contribution by citing it. Therefore, the scientific impact of a paper depends on the

[1] "Athena" R.C., Artemidos 6 & Epidavrou 15125, Marousi, Greece
E-mail: {ilias.kanellos,vergoulis}@imis.athena-innovation.gr
[2] NTU Athens, 9, Iroon Polytechniou St, 15780, Athens, Greece
[3] TU Wien, Favoritenstraße 9-11/194-04, 1040 Vienna, Austria
E-mail: dimitris@ec.tuwien.ac.at

network of citations. In this work, we focus on the *short-term impact* (STI) of a paper, quantified by the number of citations it acquires in the near future (referred to as "new citations" [30], or "future citation counts" [13]). Specifically, we address the research problem of ranking papers via their expected short-term impact.

Existing work typically assigns to each paper a proxy score estimating its expected short-term impact. These scores are determined by a stochastic process, akin to PageRank [24], modelling the impact flow in the citation network. The important concern here is to account for the *age bias* inherent in citation networks [6, 14, 39, 20]: as papers can only cite past work, recent publications are at a disadvantage having less opportunity to accumulate citations. A popular way to address this is by introducing time-awareness into the stochastic process, by favoring either recent papers or recent citations [30, 25, 13]. Nonetheless, it has been shown [16] that these methods still leave enough space for further improvements.

In this paper, we argue that there is an additional, previously unexplored mechanism that governs where future citations end up. We posit that recent citations strongly influence the short-term impact, in that the level of *attention* papers currently enjoy will not change significantly in the very near future. We investigate this hypothesis and find that it holds to a certain degree across different citation networks. Hence, we introduce an attention-based mechanism, reminiscent of a time-restricted version of preferential attachment [2], that models the fact that recently cited papers continue getting cited in the short-term.

The proposed paper ranking method, called AttRank, describes an iterative process simulating a researcher reading existing literature. At each step in the process, the researcher has studied some paper and decides what to read next among three options: (a) pick a reference from the current paper, (b) pick a recent paper, and (c) pick a currently popular paper. The first option models the impact flow of impact from citations, the second option mitigates age bias, while the third option models the aforementioned attention-based mechanism of network growth. We can guarantee that, if the probabilities are properly configured, this process will always converge (see Theorem 1). This converged AttRank score of each paper acts as a proxy to its unknown short-term impact. Hence, to estimate their STI ranking we rank papers in decreasing order of their AttRank score.

To evaluate AttRank's effectiveness in identifying papers with high short-term impact, we perform an extensive experimental evaluation, on four citation networks from various scientific disciplines. We measure effectiveness as the ranking accuracy with respect to the ground truth STI ranking. We investigate the importance of the attention mechanism in achieving high effectiveness. We also compare AttRank against several state-of-the-art methods, which are carefully tuned for each experimental setting. Our findings indicate that across almost all settings, AttRank outperforms prior work.

The contributions of this work are as follows:

- We study the problem of ranking papers by their short-term impact (STI), and observe that among the top ranking papers we not only find papers published recently, but also papers that have just recently become popular.
- We propose a popularity-based model of growth for the citation network that seeks to explain the aforementioned observation. We then introduce paper ranking method, called AttRank, that materializes this model.

- We perform an extensive experimental evaluation that highlights the importance of the popularity-based growth mechanism in achieving superior performance against state-of-the-art methods. Specifically, we find that AttRank achieves higher positive correlations to STI rankings, by up to 0.077 units compared to its competitors, and higher nDCG values, by up to 0.098 units.
- AttRank's implementation is scalable and can be executed on very large citation networks. It will be made publicly available under the GNU/GPL license.

Outline. The remaining of this paper is structured as follows. In Section 2, we introduce the problem and related concepts. In Section 3, we investigate the attention-based mechanism and introduce our method, AttRank. Then, in Section 4, we experimentally evaluate AttRank's effectiveness in comparison to the state-of-the-art paper ranking methods. In Section 5, we discuss related work. We conclude our contribution in Section 6.

2 Background

Citation Network. We represent a collection P of papers as a directed graph, which we call the *citation network*. Each node p_i in this graph corresponds to a paper, while each directed edge from p_j to p_i corresponds to a reference from paper p_j to paper p_i.

A citation network can be represented by its *citation matrix* $\mathbf{C}$, where $C[i,j] = 1$, iff paper p_j cites paper p_i, or $C[i,j] = 0$, otherwise. We denote as t_{p_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 popular centrality metrics for citation networks.

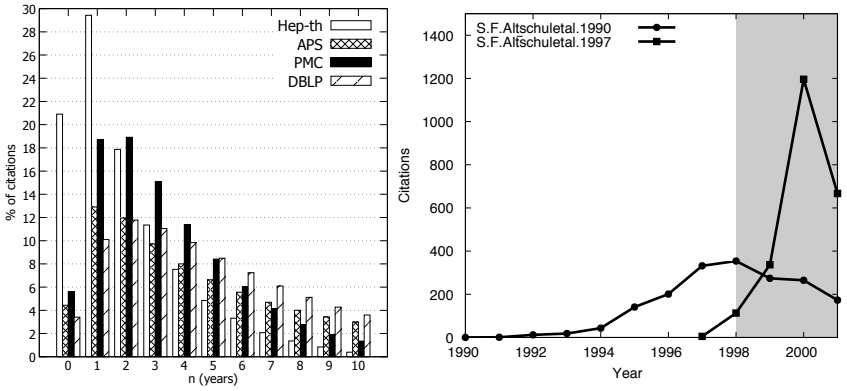
Citation Count. The *citation count* (CC) of a paper p_i is the in-degree of its corresponding node, computed as $CC(p_i) = \sum_j C[i,j]$.

PageRank. PageRank [24] measures the importance of a node in a network, by defining a random walk with jumps process. In the context of citation networks, the process simulates a "random researcher", who starts her work by reading a paper. Then, with probability $\alpha \in [0,1]$, she picks another paper to read from the reference list, or, with probability $1-\alpha$, chooses any other paper in the network at random. The PageRank score of a paper p_i indicates the probability of a random researcher reading it, and satisfies:

$$PR(p_i) = \alpha \cdot \left(\sum_j S[i,j] \cdot PR(p_j) \right) + (1-\alpha) \cdot \left(\frac{1}{|P|} \right), \tag{1}$$

where $\mathbf{S}$ is a stochastic matrix derived from the citation matrix as follows. Let k_i denote the number of papers referenced by p_i. Then, $S[i,j] = \frac{1}{k_j}$, iff paper p_j cites paper p_i, $S[i,j] = 0$, iff p_j does not cite p_i but cites at least one other paper, and $S[i,j] = \frac{1}{|P|}$, iff paper p_j cites no paper (i.e., is a dangling node).

Short-Term Scientific Impact (STI). 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 [14,6,39]. This is due to the inherent characteristics of citation networks: the references of a



(a) (b)

Fig. 1: (a) Empirical distribution of the fraction of total citations received by papers n years after their publication ($n \leq 10$), from four citation networks (see Section 4.1) (b) A comparative yearly citation count of two papers.

paper are fixed, and there is a delay between a paper's publication and its first citation, known as *citation lag* [9]. This phenomenon is best portrayed in Figure 1a, where it is shown that, for different citation networks (introduced in Section 4.1), the bulk of citations comes a few years after the paper is published. In contrast, the *short-term impact* [16], also called the number/count of new/future citations [30,13], of a paper looks into a future state of the citation network and reflects the level of attention (in terms of citations) a paper will receive in the near future.

As a motivating example for the importance of short-term impact, examine the case of two seminal papers in the bioinformatics literature. The first, published in 1990, introduces the initial version of the popular BLAST alignment algorithm, while the second, published in 1997, presents an improved alignment algorithm by the same team. Figure 1b comparatively illustrates their yearly citation counts.[1] Now, consider a bioinformatics researcher living in the year 1998. At that point in time, the older paper has a higher citation count. However, the newer paper is clearly more popular as it has a greater short-term impact, evidenced by the number of citations it collects in the next three years (highlighted in the figure). The 1998 researcher would benefit from being able to identify the newer paper as potentially having a higher short-term impact.

Let $P(t)$ denote the subset of papers published until time t, and let $\mathbf{C}(t)$ denote the state of the citation matrix at t, i.e., including only citations from papers in $P(t)$. Note that just the *content*, and not the shape, of $\mathbf{C}(t)$ changes with t; $\mathbf{C}(t)$ contains rows and columns for all papers in P (even those published after time t). Given a time *horizon* of τ units, the short-term impact at current time t_N of a paper $p_i \in P(t_N)$ is defined as:

$$STI(p_i; t_N, \tau) = \sum_j \left(C(t_N+\tau)[i,j] - C(t_N)[i,j]\right).$$

[1] Based on open citation data from COCI (`http://opencitations.net/download`).

Table 1: Recently popular papers in top-100 (accord. to STI)

Dataset	hep-th	APS	PMC	DBLP
Recently Popular	41	54	54	63

As $\mathbf{C}(t_N+\tau) - \mathbf{C}(t_N)$ contains a non-zero entry for citations made during the short-term period $[t_N, t_N+\tau]$, STI essentially counts the number of citations a paper receives in this period.

Some observations are in order. First, the time horizon τ is a user-defined parameter that specifies how long in the future one should wait for citations to accumulate. An appropriate value may depend on the typical duration of the research cycle (preparation, peer-reviewing, and publishing) specific to each scientific discipline. Second, it is important to emphasize that STI can only be computed in retrospect; at current time t_N, the future citations are not yet observed. Thus, any method that seeks to identify papers with high STI has to employ a mechanism to account for the unobserved future citations.

With these remarks in mind and similar to prior work [30, 13, 16], we study the following problem.

Problem 1 Given the state $\mathbf{C}(t_N)$ of the citation network at current time t_N, return a ranking of papers in $P(t_N)$ such that it matches their ranking by short-term impact $STI(\cdot; t_N, \tau)$ for a given time horizon τ.

3 Our Approach

The main hypothesis of our work is that researchers tend to read and cite trending papers, i.e., papers that have recently received significant attention from the scientific community. To investigate this hypothesis, we explore four citation networks (as per the default experimental configuration discussed in Section 4.1), and count how many top-100 papers were recently popular, based on STI (i.e., were among the top cited in the past 5 years). As we see in Table 1, roughly half of the top-100 papers were, indeed, recently popular.

This observation validates our assumption that the level of attention a paper has recently attracted is indicative of its ability to attract citations in the short-term. We thus propose the following method to quantify recent attention. Let $\mathbf{C}[t_N - y : t_N]$ denote the citation matrix taking into account only citations made during the past y years. The recent attention, simply called *attention*, of a paper p_i is calculated as:

$$A(p_i) = \frac{\sum_j C[t_N - y : t_N][i,j]}{\sum_i \sum_j C[t_N - y : t_N][i,j]}, \tag{2}$$

which corresponds to the fraction of total citations in the last y years, which paper p_i received. Hyperparameter y induces time-awareness, allowing us to capture the most recent dynamics; its value is to be tuned for the specific ranking problem.

Attention, however, is not the only mechanism that governs which papers researchers read. Naturally, researchers may read a paper cited in the reference list of another paper. Moreover, similar to previous work [30, 25], we assume that researchers also read recently published papers. Specifically, we capture the recency

of a paper p_i using a score that decays exponentially based on the paper's age:

$$T(p_i) = c \cdot e^{w \cdot (t_N - t_{p_i})}, \tag{3}$$

where t_N is the current time, t_{p_i} denotes the publication time of paper p_i, hyperparameter w is a negative constant (as $t_N - t_{p_i} \geq 0$), and c is normalization constant so that $\sum_i T(p_i) = 1$. To calculate a proper w value, a similar procedure like the one used in [25] can be followed (see also Section 4.2).

Combining these mechanisms, we assume that a researcher may read a paper for one of the following reasons: the paper gathered attention recently, the paper was recently published, or the paper was found in another paper's reference list. We model this behavior with the following random process. A researcher, after reading paper p_i, chooses to read any other paper from p_i's reference list, with probability α. With probability β she chooses a paper based on its attention. This behavior essentially makes recently rich papers even richer, and is reminiscent of a time-restricted preferential attachment mechanism of the Barabási-Albert model of network growth [2]. Finally, with probability γ she chooses any paper with a preference towards recently published ones.

Specifically, our model computes a score $AR(p_i)$, called AttRank, for each paper p_i that satisfies the following recurrence:

$$AR(p_i) = \alpha \cdot \left(\sum_j S[i,j] \cdot AR(p_j) \right) + \beta \cdot A(p_i) + \gamma \cdot T(p_i), \tag{4}$$

where the coefficients $\alpha, \beta, \gamma \in [0,1]$ and $\alpha + \beta + \gamma = 1$, and $\mathbf{S}$ is the stochastic citation matrix as defined in PageRank. As is typical, the optimal values for the three coefficients are to be determined by validation.

Two special values for coefficient β are worth mentioning. First, observe that when $\beta = 0$, a setting we call NO-ATT (for no attention), the model becomes similar to time-aware methods that address the inherent bias against new papers in citation networks (see also Section 5, and [16] for a thorough coverage of such approaches). Note that additionally setting $w = 0$ in Eq. 3 recovers PageRank. Second, when $\beta = 1$, a setting we call ATT-ONLY (for attention only), AttRank is solely based on the attention mechanism, assuming that the recent citation patterns will persist exactly in the near future. To the best of our knowledge, ATT-ONLY has not been considered in the literature as a means to estimate the short-term impact of a paper. As we show in Section 4, attention alone is a powerful mechanism, often more effective than existing approaches. However, $\beta = 1$ is never the optimal setting; it is always better to consider attention in combination with the other two citation mechanisms.

Equation 4 describes an iterative process for computing the AR vector: starting with a random value, at each step update the vector with the right hand side of Eq. 4. This process is repeated until the AR values converge. The following theorem, ensures that convergence is achievable.

Theorem 1 *The iterative process defined by Eq. 4 converges.*

Proof We can rewrite Equation 2 in matrix form as:

$$AR = \mathbf{R}\, AR \tag{5}$$

where $\mathbf{R}$ is a matrix satisfying:

$$R[i,j] = \alpha \cdot S[i,j] + \beta \cdot A(p_i) + \gamma \cdot T(p_i) \tag{6}$$

In other words, matrix $\mathbf{R}$ is a modified citation matrix, artificially expanded with directed edges from any node to any other in the network. For each column c of matrix $\mathbf{R}$, the following property holds:

$$\begin{aligned} \sum_i R[i,c] &= \alpha \cdot \sum_i S[i,c] + \beta \cdot \sum_i A(p_i) + \gamma \cdot \sum_i T(p_i) \\ &= \alpha + \beta + \gamma = 1 \end{aligned} \tag{7}$$

Equation 7 holds, because (a) the sum over all A scores equals 1, since these scores are probabilities, (b) by definition the sum over all T scores, equals 1, and (c) $\mathbf{S}$ is a column stochastic matrix, thus by definition each of its columns has values that sum up to 1. Hence, matrix $\mathbf{R}$ is also a stochastic matrix and Equation 5 represents a power method equation applied on matrix $\mathbf{R}$.

The power method on a stochastic matrix converges, under the following conditions [18]: matrix $\mathbf{R}$ must be irreducible and aperiodic. Irreducibility is guaranteed for any matrix that corresponds to a strongly connected graph, while aperiodicity is guaranteed if $R[i,i] > 0$. Both conditions hold for matrix $\mathbf{R}$, due to the fact that we added artificial edges connecting each node to any other node in the citation network, as well as itself.[2] Therefore, the iterative process is guaranteed to converge to a single vector of AttRank scores. □

4 Evaluation

This section presents a thorough experimental evaluation of our approach for ranking papers based on their short-term impact. Specifically, Section 4.1 discusses the experimental setup and evaluation approach taken. Section 4.2 investigates the effectiveness of our proposed method and the importance of the attention-based mechanism. Section 4.3 compares AttRank with existing approaches from the literature. Finally, Section 4.4 discusses the convergence rate of AttRank.

4.1 Experimental Setup

Datasets. We consider four datasets in our experiments:

1. arXiv's high energy physics (hep-th) collection, which was provided by the 2003 KDD cup.[3] This collection consists of approximately 27,000 papers with 350,000 references, written by 12,000 authors from 1992 to 2003.
2. A collection of papers provided by the American Physical Society (APS)[4], which contains about 500,000 papers with 6 million references, written by about 389,000 authors from 1893 to 2014.

[2] Since the time-based vector is exponentially decreasing based on paper age, $T(p_i) > 0, \forall p_i$. Thus, in matrix $\mathbf{R}$ there is a link (however low-weight) from all nodes to all others.

[3] `http://www.cs.cornell.edu/projects/kddcup/datasets.html`

[4] `https://journals.aps.org/about`

Table 2: Correspondence of the Test Ratio to the Time Horizon

Test Ratio	Time Horizon τ (in years) hep-th	APS	PMC	DBLP
1.2	1	4	1	1
1.4	2	7	2	3
1.6	3	10	2	4
1.8	4	13	3	6
2.0	5	16	3	7

3. A collection of open access papers from pubmed central[5] (PMC), which consists of about 1 million papers with 665,000 references, written by 5 million authors, from 1896 to 2016.
4. A collection of about 3 million papers and 25 million references, written by more than 1.7 million authors, from the computer science domain (DBLP)[6], published from 1936 to 2018.

Evaluation Methodology. To evaluate the effectiveness of AttRank in ranking papers based on their short-term impact, we construct a *current* and a *future* state of the citation network. We partition each dataset according to time in two parts, each having equal number of papers. We use the older half to construct the current state of the citation network, denoted as $\mathbf{C}(t_N)$. All ranking methods will be based on this network acting as the "training" subset. We use parts of the newer half to construct the future state of the network, denoted as $\mathbf{C}(t_N+\tau)$. All ranking methods will be evaluated based on this network acting as the "test" subset.

Specifically, the future state is constructed as follows. We vary the size, in terms of number of papers, of the future state relative to the size of the current state. Thus we do not vary the time horizon τ directly, but rather the *test ratio*, which is the relative size of the future with respect to the current network. We consider values for the test ratio among $\{1.2, 1.4, 1.6, 1.8, 2.0\}$, where 2.0 corresponds to using all citations in the dataset to define the future state. In some experiments we fix the test ratio to a default value of 1.6, meaning that the future state contains 30% more papers than the current state. Table 2 presents, for each dataset, the length in years of the time horizon that corresponds to each test ratio value. Note, that the relationship between test ratio and τ is not linear, due to the non-constant number of published papers per year and the fact that most datasets contain incomplete entries for the last year they include.

Given the future state of the citation network, we can compute the STI of each paper as per its definition (see Section 2). Similar to previous approaches in the literature [25, 30, 13, 16], the ranking of papers based on their STI forms the *ground truth*. Any paper ranking method is oblivious of the future state $\mathbf{C}(t_N+\tau)$ of the citation network, and hence the ground truth, and only uses the current state $\mathbf{C}(t_N)$ to derive a ranking. To quantify the effectiveness of a method, we compare its produced ranking to the ground truth, using the following two measures:

– Spearman's ρ [28] is a non-parametric measure of rank correlation. It is based on the $L1$ distance of the ranks of items in two ranked lists and provides a

[5] `https://www.ncbi.nlm.nih.gov/pmc/`
[6] `https://aminer.org/citation`

Table 3: AttRank's parameterization space.

Parameter	min	max	step
α	0.0	0.5	0.1
β	0.0	1.0	0.1
γ	0.0	0.9	0.1
y	1	5	1

quantitative measure to compare how similar these lists are. Its values range from -1 to 1 with 1 denoting perfect correlation, -1 denoting perfect negative correlation and 0 denoting no correlation.

– Normalized Discounted Cumulative Gain at rank k (nDCG@k) is a rank-order sensitive metric. 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.e., the short-term impact, of the paper that appears at the i-th position on the method's ranking. The nDCG@k is the paper's DCG divided by the ideal DCG, achieved when the method's ranking coincides with the ground truth. In our evaluation, we consider values of k among $\{5, 10, 50, 100, 500\}$, with $k = 50$ being used as a default value.

Spearman's ρ calculates an overall similarity of the given ranking with the ground truth ranking. In contrast, nDCG@k measures the agreement of the two rankings on the top-ranking papers.

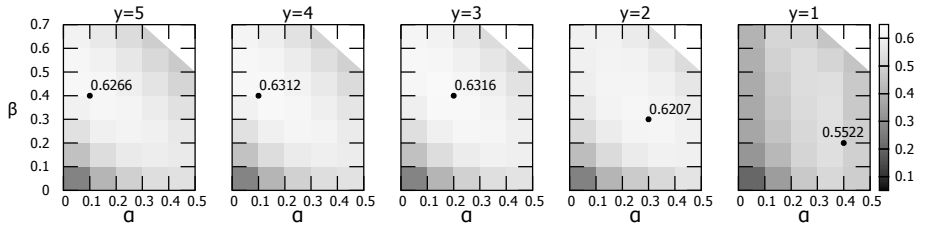
4.2 Ranking Effectiveness

In this section, we investigate AttRank's effectiveness for the default experimental setting (test ratio equal to 1.6), while varying its parameters, α, β, γ, and the number y of past years used to calculate the attention of a paper. The range of values tested are shown in Table 3. For each metric, we discuss AttRank's parameterization that achieves the best ranking effectiveness.

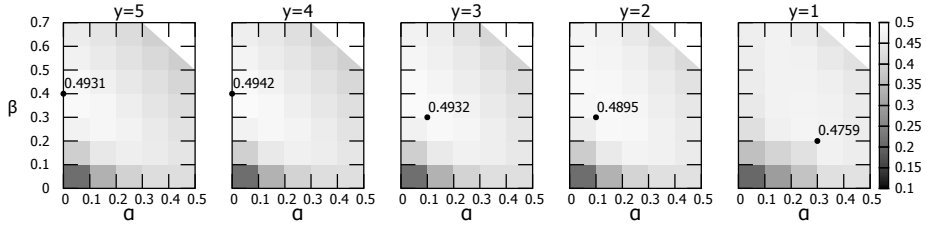
First, however, we discuss how we set the value of the exponential factor w of Equation 3. We follow a similar approach as the one used in [25]. For each dataset, we use an exponential function of the form $e^{\widetilde{w}y}$, to fit the tail of the distribution of the random variable Y that models the probability of an article being cited n years after its publication. Figure 1a illustrates the empirical probability distribution for each dataset. The factor $\widetilde{w}$ of the fitting function is used as the w value. Following this procedure, we calculate $w = -0.48$ for hep-th, $w = -0.12$ for APS and $w = -0.16$ for PMC and DBLP.

4.2.1 Effectiveness in terms of Correlation

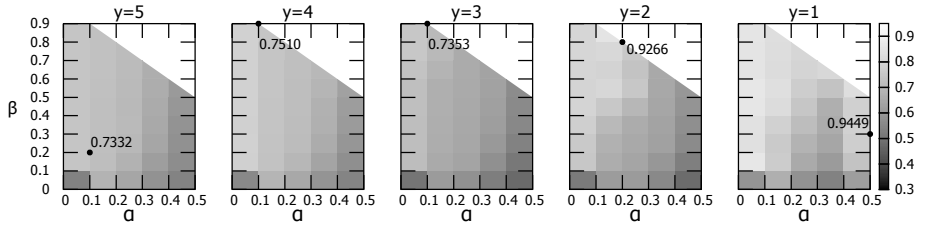
In this experiment, we measure the ranking effectiveness of AttRank, in terms of Spearman's ρ to the ground truth ranking by STI. We can visualize the effectiveness of each tested parameterization as a heatmap over the α–β space for different values of y. Indicatively, we show the heatmaps, for the various parameter settings on DBLP and PMC in in Figures 2a and 2b, respectively (results on the other datasets are similar and the corresponding heatmaps can be found in Appendix A). The heatmaps show the results varying parameters α, and β; parameter



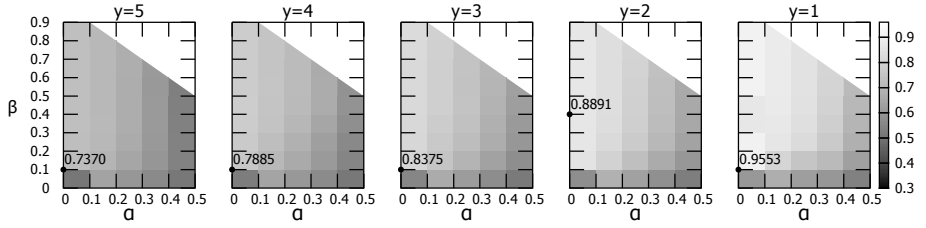
(a) Correlation on the DBLP dataset.



(b) Correlation on the PMC dataset.



(c) nDCG@50 on the DBLP dataset.



(d) nDCG@50 on the PMC dataset.

Fig. 2: Heatmaps depicting the effect of the parameterization of AttRank to its effectiveness in terms of the correlation and nDCG@50 metrics, for the DBLP and PMC datasets. The best value achieved for each metric is also depicted.

γ is implied since $\alpha + \beta + \gamma = 1$. We expect that as α increases, AttRank simulates researchers that predominantly prefer reading papers from reference lists and rarely choose papers based on their age, or on whether they have been recently popular. Thus, as α increases, AttRank gradually reduces to simple PageRank, with a small probability of random jumps. Since references are made only to papers published in the past, researchers increasingly arrive at older papers when following references with high probability. As a result, for large α values, AttRank promotes older papers and, thus, its correlation to the ground truth is expected

to drop. Most importantly, the heatmaps validate the role of the attention scores, since for $\beta = 0$ (NO-ATT) we observe significantly lower correlations (notice the darker color on the bottom left corner of the heatmaps). Similarly, lower correlations were observed when $\beta = 1$ (ATT-ONLY).

From the produced correlation scores, we firstly gather that AttRank correlates at least moderately to the ground truth ranking for all datasets in its best setting (i.e., $\rho > 0.49$). Further, we observe that the optimal value for the number of past years y, used to calculate the attention score is $y = 3$ and $y = 4$ on DBLP and PMC, respectively, while it's $y = 3$ on APS, and $y = 1$ on hep-th (see Appendix A).

Interestingly, the first three datasets follow relatively similar citation patterns (see Figure 1a), with papers having a citation peak at $2 - 3$ years after their publication, while hep-th shows quicker citation peaks. Intuitively then, it makes sense to use a smaller value of y to calculate attention scores for hep-th. Since its research trends may change faster, a larger time window to calculate the attention would reflect past research trends, and not current ones. On APS, PMC, and DBLP, in contrast, papers gather citations at a slower rate, thus a larger y value is more likely to reflect current research preferences.

Based on these experiments, we also identify the optimal parameterization that achieves maximum correlation per dataset. We find the best settings of $\{\alpha,\beta,\gamma,y\}$ at $\{0.3, 0.4, 0.3, 1\}$ for hep-th ($\rho = 0.6519$), $\{0.3, 0.3, 0.4, 3\}$ for APS ($\rho = 0.6295$), $\{0.0, 0.4, 0.6, 4\}$ for PMC ($\rho = 0.494$), and $\{0.2, 0.4, 0.4, 3\}$ for DBLP ($\rho = 0.6316$). To illustrate the significance of the attention mechanism, compare these results to the maximum values for $\beta = 0$ (NO-ATT). These are 0.56, 0.581, 0.411, and 0.529 for hep-th, APS, PMC, and DBLP, respectively. Accordingly, for $\beta = 1$, these values are 0.615, 0.537, 0.45, 0.571.

4.2.2 Effectiveness in terms of nDCG@50

We repeat the effectiveness analysis, this time considering the nDCG@50 metric. Indicatively, we present the heatmaps for DBLP and PMC in Figures 2c and 2d, respectively (results are similar on the other datasets and can be found in Appendix A). An interesting observation is that with regards to only capturing the papers with the highest short-term impact, smaller time windows on which the attention scores are calculated seem to be more suitable. We observe that as y increases, the overall nDCG values decrease fast (notice the darker hues when $y > 1$). We expect that by further increasing y, the nDCG would further drop. This is because by increasing the time window on which we calculate the attention we, re-introduce the inherent age bias of citation networks, and the papers with the highest attention scores no longer reflect current research trends. The same observation holds for increased values of α when $y > 1$. As α increases, the PageRank component dominates AttRank, giving advantage to older papers that are not necessarily at the current focal point of research. This observation is evident from the darker hues on the heatmaps for values of α close to 0.5.

Finally, we determine the parameterization that achieves the best nDCG@50 per dataset. We find the best settings of parameters $\{\alpha,\beta,\gamma,y\}$ at $\{0.0,0.4,0.6,1\}$ for hep-th ($nDCG = 0.8930$), $\{0.3,0.5,0.2,3\}$ for APS ($nDCG = 0.7293$), $\{0.0,0.1,0.9, 1\}$ for PMC ($nDCG = 0.9553$) and $\{0.5, 0.3, 0.2, 1\}$ for DBLP ($nDCG = 0.9449$). As before, we observe that the attention vector plays a non-negligible role in achieving the maximum nDCG on all datasets (i.e., $\beta > 0$). Indicatively, the

maximum nDCG@50 values for $\beta = 0$ are 0.669, 0.635, 0.6, and 0.663 for hep-th, APS, PMC, and DBLP, respectively. Accordingly, for $\beta = 1$ these values are 0.89, 0.692, 0.916, 0.916.

4.3 Comparative Evaluation

In this section, we compare AttRank to existing approaches for impact-based paper ranking. Based on a recent experimental evaluation [16], we select the five methods found to be most effective in ranking by short-term impact.

- **CiteRank (CR)**. This PageRank-based method calculates the "traffic" towards papers by researchers that prefer reading recently published papers when performing random jumps [30]. It uses parameters $\alpha \in (0, 1)$ and $\tau_{dir} \in (0, \infty)$, where α models the probability with which researchers follow references from papers they read and τ_{dir} models an aging factor, which determines the papers which random researchers are more likely to select when performing random jumps. In the original work, their optimal settings are found for $\{\alpha, \tau_{dir}\}$ set to $\{0.48, 1\}$, $\{0.5, 2.6\}$,$\{0.31, 1.6\}$,$\{0.55, 8\}$.
- **FutureRank (FR)**. This method is based on PageRank and HITS [17]. It applies mutual reinforcement from papers to authors and vice versa, while additionally using time-based weights to promote recently published papers [25]. It uses four parameters: $\alpha, \beta, \gamma \in (0, 1)$, and $\rho \in (-\infty, 0)$. Parameter α is taken from PageRank, β is the coefficient of an author-based score vector, and γ is the coefficient of time-based weights. These weights depend on ρ, which modifies an exponentially decreasing function. In the original work the optimal settings of $\{\alpha, \beta, \gamma, \rho\}$ are $\{0.4, 0.1, 0.5, -0.62\}$, and $\{0.19, 0.02, 0.79, -0.62\}$.
- **Retained Adjacency Matrix (RAM).** This citation count variant uses a citation age-weighted adjacency matrix [13]. It uses a parameter $\gamma \in (0, 1)$ as the base of an exponential function, to modify citation weights, based on their age. The authors find $\gamma \in \{0.3, 0.6, 0.71\}$ as as the optimal settings.
- **Effective Contagion Matrix (ECM).** This method, based on Katz centrality, operates over a citation age-weighted adjacency matrix [13] and calculates weights of citation chains. It uses parameters, $\alpha, \gamma \in (0, 1)$, where γ is taken from RAM, and α is used to decrese citation chain weights as they increase in length. In the original work, the authors find the best settings of $\{\alpha, \gamma\}$ to be $\{0.1, 0.3\}$ or $\{0.007, 0.71\}$.
- **WSDM cup's 2016 winner (WSDM)**. We consider the winning solution [11] of a scholarly article ranking challenge. This method uses three bipartite networks (papers-authors, papers-papers, and papers-venues). It calculates paper scores by aggregating scores propagated to papers by other papers, by their authors, and their venues, additionally using scores based on paper in- and out-degrees. Paper scores are calculated iteratively, based on a fixed small number of iterations. The method uses parameters $\alpha, \beta \in \mathbb{R}$, as coefficients of each paper's in- and out-degree, to calculate paper scores, and the number of iterations, i. The authors use $\{\alpha, \beta\} = \{1.7, 3\}$ in their work and set $i \in \{4, 5\}$.

The optimal parameterization for the competitors is presented in each work. However, these suggested values result from the use of particular datasets and specific experimental settings, which differ among works. Therefore, in our evaluation,

Table 4: Parameterization space of competitors.

Method	Parameter	min	max	step
CR	α	0.1	0.7	0.2
	τ_{dir}	2	10	2
FR	α	0.1	0.5	0.1
	β	0.0	0.9	0.1
	γ	0.0	0.9	0.1
	ρ	-0.82	-0.42	0.2
RAM	γ	0.1	0.9	0.1
ECM	α	0.1	0.5	0.1
	γ	0.1	0.5	0.1
WSDM	α	1.1	2.3	0.3
	β	1	5	1
	i	4	5	1

we extensively tuned all competitors, to ensure a fair comparison of their effectiveness in ranking based on STI. Table 4 presents the examined parameter sets.[7] In total, we used 20 different settings for CR, 120 settings for FR, 9 settings for RAM, 25 settings for ECM, and 50 settings for WSDM. Note, that since WSDM requires venue data, we ran this method only on the PMC and DBLP datasets, for which this data was available. Further, we ran all iterative methods until the convergence error drops below 10^{-12}, to ensure that all scores approach their final values and further iterations are not expected to change the ranking of papers.

In addition to these existing approaches, we also consider two variants of AttRank that better demonstrate the effect of the attention mechanism. The first, denoted as NO-ATT, removes the attention mechanism in AttRank, i.e., sets $\beta = 0$. Conversely, the second, denoted ATT-ONLY, considers only the attention mechanism in AttRank, i.e., sets $\beta = 1$.

4.3.1 Effectiveness in terms of Correlation

In this experiment, we measure the correlation of each method's ranking to that of the ground truth. We vary the test ratio of the size of networks according to Section 4.1. For each dataset and test ratio, we choose the parameterization with the best correlation. Figure 3 presents the results.

We observe that AttRank's ranking better correlates to the ground truth ranking, compared to all competitors on all settings for the hep-th, APS, and DBLP datasets. In particular, AttRank increases correlation by up to 0.055 units on hep-th, by up to 0.057 on APS, and by up to 0.077 on DBLP with respect to the best competitor. Further, on most settings AttRank correlates better to the ground truth ranking on PMC, by up to 0.027, compared to the best competitor, while marginally losing to FR on two settings (the correlation values observed differ by < 0.01). It is worth highlighting that FR achieves such a good correlation only for PMC; in the other datasets, it is outperformed by other existing methods. In

[7] The settings were chosen so as to include, for each parameter, values close, or equal, to those suggested in the original works. Note, that since the value of one parameter may restrict the range of others, the total number of settings does not equal the sum of all individual parameter settings. Note also, that some works do not provide a formal proof of convergence. Hence, we exclude the parameter ranges in Table 4 which resulted in non-convergence.

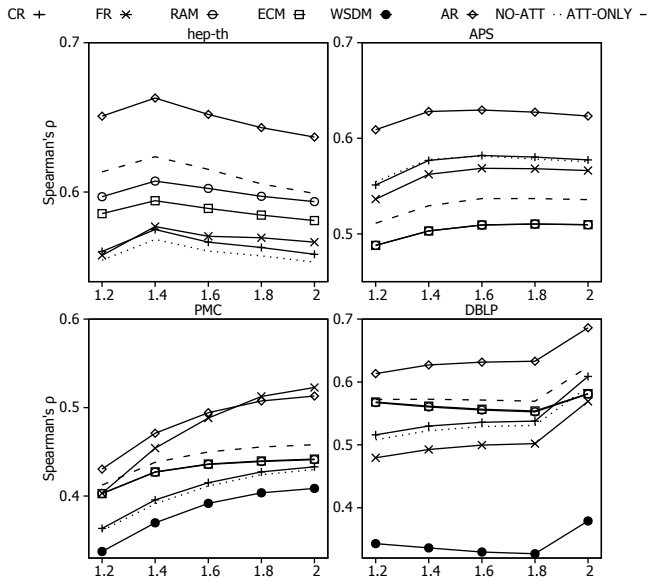


Fig. 3: Effectiveness of all methods in terms of correlation. The x-axis represents the varying test ratio.

contrast, AttRank is robust across datasets and settings with a large correlation gain over all competitors (except FR in PMC).

Our method's performance can be attributed to the fact that, compared to the time-aware competitors, it does not simply promote papers recently cited, or recently published. Instead, because of the attention mechanism, it heavily promotes well-cited, recent papers, compared to lesser cited recent papers. As discussed in Section 3, recently popular papers indeed remain popular. Moreover, our method promotes older papers that are still heavily cited. The importance of the attention mechanism is illustrated by the fact that in two datasets ATT-ONLY outperforms existing methods. Turning off attention completely, i.e, the NO-ATT method, results in subpar performance, except in one dataset. Most importantly, in all cases, the effectiveness is increased when the attention mechanism is balanced with the other mechanisms in AttRank.

4.3.2 Effectiveness in terms of nDCG

In this section, we measure the nDCG achieved by each method with regards to the ground truth. We conduct two experiments: in the first, we set $k = 50$ as the cut-off rank when computing nDCG, varying the test ratio. In the second experiment we use the default test ratio (at 1.6) and measure nDCG varying k.

Figure 4 presents the results varying the test ratio. For each setting, we select the parameterization of each method that gives the best nDCG@50 value. In general, as we look further into the future, i.e., increase the test ratio, the ranking accuracy of all methods drops; the effect is more pronounced in the APS dataset, and less in hep-th. In all cases, AttRank outperforms all competitors, with the

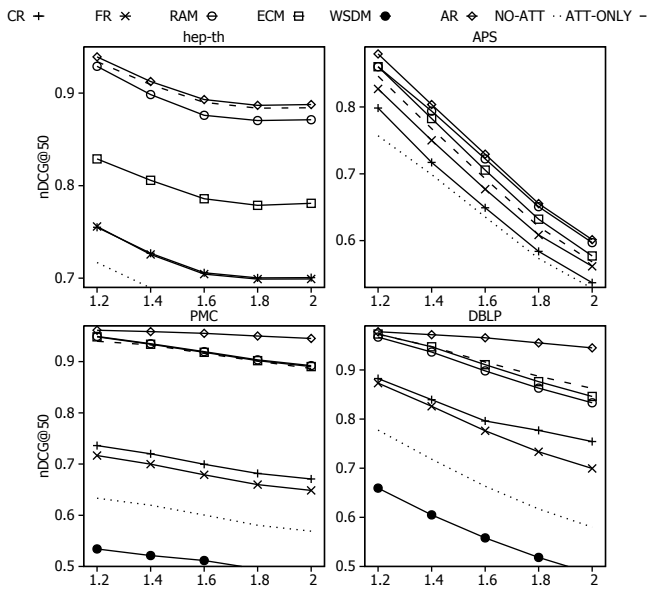


Fig. 4: Effectiveness of all methods in terms of nDCG@50. The x-axis represents the varying test ratio.

margin increasing with the test ratio in two datasets. In particular, our method improves nDCG@50 by up to 0.017 units on hepth, 0.018 on APS, 0.053 on PMC, and 0.098 on DBLP, compared to the best existing method. It is worth mentioning that the best existing method varies across datasets, being either RAM or ECM.

Figure 5 presents the results varying k for the a test ratio of 1.6. For each setting, we select the parameterization of each method that gives the best nDCG@k value. In general we observe that AttRank is at least on par, and mostly outperforms all rivals on all datasets, with the sole exception of nDCG@5 on APS (the measured difference compared to the best competitor is 0.015). Specifically, AttRank achieves a higher nDCG value of up to 0.017 units on hep-th, up to 0.013 units on APS (except nDCG@5), up to 0.035 on PMC, and up to 0.111 units on DBLP. Additionally, for small values of k ($k = \{5, 10\}$) AttRank achieves nDCG values close to 1 on three out of four datasets (hep-th, PMC, and DBLP). The best competitors are again RAM and ECM, depending on the dataset.

Regarding the special cases of AttRank, we observe in both Figures 4 and 5, that excluding attention (NO-ATT) results in a significant drop in nDCG. On the other hand, attention alone (ATT-ONLY) outperforms most existing methods except in APS. As also observed in the case of Figure 3, carefully balancing the mechanisms in AttRank leads to a considerable improvement in ranking accuracy.

4.4 Convergence of AttRank

AttRank involves an iterative process, similar to PageRank, to compute scores for papers. Specifically, we can view AttRank as a PageRank variant, where the

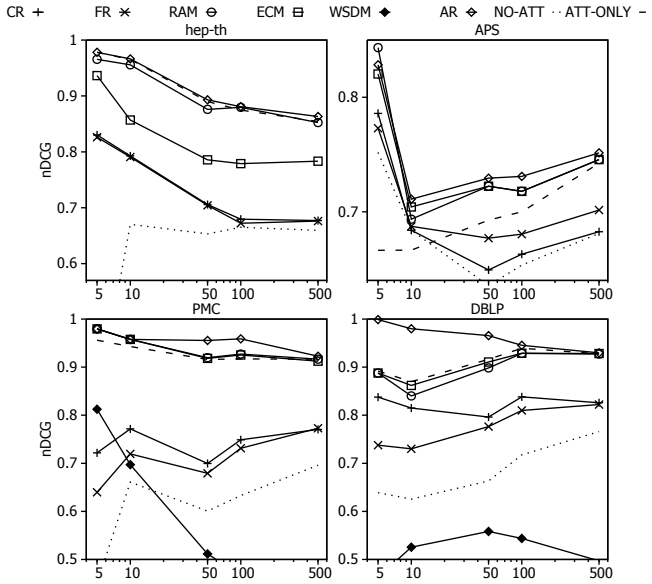


Fig. 5: Effectiveness of all methods in terms of nDCG@k on the default test ratio. The x-axis represents the varying value of k.

random jump vector is replaced by two vectors, the attention-based vector and the publication age-based vector, and thus PageRank's random jump probability $1-\alpha$ is divided among β, γ in AttRank. The convergence of AttRank is thus affected by the same factors as PageRank's; an in-depth discussion of PageRank's convergence properties can be found in [18]. The most important property is that as $\alpha \to 1$, the convergence rate decreases and more iterations are required.

Following the discussion in Section 4, however, large values of parameter α do not favor ranking based on short-term impact, and AttRank's optimal effectiveness is always achieved for $\alpha \leq 0.5$. Additionally, as $\alpha \to 0$, AttRank tends to depend increasingly on the sum of the attention- and time-based vectors. Thus, the number of iterations required for convergence decreases, with the limit case $\alpha = 0$ requiring a single iteration (i.e., the calculation of the attention- and time-based vectors).

Overall, AttRank is expected to converge faster than PageRank and other variants (PageRank has been used with $\alpha = 0.5$ on citation networks [6,21]). In our experiments, AttRank converges in less than 30 iterations for hep-th, APS, and DBLP, and less than 20 iterations for PMC, for $\alpha = 0.5$ and a convergence error of $\epsilon \leq 10^{-12}$, with the number of iterations decreasing for smaller values of α. Compare this to the maximum required iterations for CR, which are 51, 46, 26, and 47, for hep-th, APS, PMC, and DBLP, respectively, for $\alpha = 0.5$. The corresponding numbers for FR (which did not, in practice, converge under all possible settings) are 35, 30, 26, and 23, for hep-th, APS, PMC, and DBLP, respectively, and for $\alpha = 0.5$.

5 Related work

In recent years, various methods have been proposed for quantifying the scientific impact of papers. In the following, we review the most important work, focusing on methods to rank papers by their expected short-term impact. For a thorough coverage of this research area refer to [1,16].

Basic Centrality Variants. A large number of methods are PageRank adaptations tailored to better simulate the way a random researcher traverses the citation network while reading papers (e.g., [38,41,27,6]). While such approaches modify the random researcher's behaviour in intuitive ways (e.g., she prefers reading cited papers that are similar to the one she currently reads), they do not address age bias, an important intrinsic issue in citation networks.

Time-Aware Methods. To alleviate age bias, a number of time-aware methods were proposed. These methods introduce time-based weights in the various centrality metric calculations, to favor either recent publications (e.g., [39,10,30,25] or recent citations (e.g., [13]), or citations received shortly after the publication of an article (e.g., [36,40]).[8]

Although the aforementioned practice has been applied to citation count variants [36,40,13] or to Katz centrality [13], most works introduce time-awareness to PageRank adaptations. This is achieved by modifying either the adjacency matrix [39,10] and/or landing probabilities in the PageRank formula (e.g., [30,25,14, 10]). In the former case, the intuition is that the random researcher avoids following references to old papers (with respect to the current time or to the publication year of the citing paper). In the latter case, the random researcher prefers selecting new papers during random jumps.

Time-awareness is shown to improve the accuracy when ranking by short-term impact. However, it fails to differentiate among recent papers favoring all equally. In reality, some papers are fitter than others and will attract more attention. To address this issue, literature proposes using additional information besides the citation network, such as paper metadata and other networks.

Metadata. An interesting approach is to incorporate paper metadata (e.g., information about authors, venues) into the ranking method. 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 [12], or the Eigenfactor [3]. The majority of approaches in this category incorporates paper metadata in PageRank-like models, to modify citation/transition matrices (e.g., [36]), or both citation/transition matrices and random jump probabilities [14,7]. An alternative to the above approaches is presented in [39] 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.

Multiple Networks. Another way to incorporate additional information is to define iterative processes on multiple interconnected networks (e.g., author-paper, venue-paper networks) in addition to the basic citation network. We can broadly discern two approaches: the first is based on mutual reinforcement, an idea orig-

[8] Note that time-aware weights with different interpretations have been proposed in a limited number of works, in particular in [7,22].

inating from HITS [17]. Methods following this approach (e.g., [25,33]) 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). In the second approach, a single graph spanning heterogeneous nodes is used for all calculations [23,15] and scores are propagated between all types of nodes during an iterative process.

Ensemble Techniques. A popular approach for improving ranking accuracy is to consider ensembles that combine the rankings from multiple methods. The majority of the 2016 WSDM Cup[9] paper ranking methods (e.g. [11,5,34]) and their extensions (like [22]) fall in this category. They combine several types of scores like in- and out-degrees, simple and time-aware PageRank scores, metadata-based scores etc., calculated on different graphs (citation network, co-authorship network, etc). For instance, the winning solution of the cup [11] (see Section 4), combines various scores derived from in- and out-degrees with scores propagated from venues, papers, and authors.

Paper Citation Prediction. A separate line of work is concerned with modeling the arrivals of citations for *individual* papers to predict their *long-term* impact. Early approaches [37,8] model the problem as a time series prediction task. Following the seminal work of [31], subsequent works model the arrival of citations using non-homogeneous Poisson [26] or Hawkes [35] processes.

This line of work is ill-suited for ranking by short-term impact for two reasons. First, it has a different goal, predicting the citation trajectory of individual papers, and as such it optimizes for the prediction error with respect to the actual citation trajectories. Second, the training process is prone to overfitting [32], and requires a long history ($\geq$ 5 years) of observed citations for each paper. In constrast, the majority of the top ranking papers by short-term impact are recent publications. For example, in the default experimental configuration of the PMC dataset (see Section 4.1) 79% of the top-100 papers are published in the last 5 years.

Discussion. The time-awareness mechanism is not sufficient for distinguishing the short-term impact of papers. As explained, recent work focuses on using additional data sources (venues, co-authorship networks, etc.) to build better informed models. However, an important limitation of this strategy is that this data is not readily available, fragmented in different datasets, not easy to collect, integrate and clean, and is often incomplete. In contrast, our approach is to rely solely on the properties of the underlying citation network, and try to better model the process with which the network evolves.

6 Conclusion

In this work, we present AttRank, a method that effectively ranks papers based on their expected short-term impact. The key idea is to carefully utilize the recent attention a paper has received. Specifically, our method models the process of a random researcher reading papers from the literature, and incorporates an attention mechanism to identify popular papers that are likely to continue receiving

[9] The task was to rank papers based on their "query-independent importance" using information from multiple interconnected networks [29].

citations, as well as a time-based mechanism to promote recently published papers that have not yet received sufficient citations.

We studied the effectiveness of our approach in terms of Spearman's rank correlation and nDCG compared to the ground truth rankings compiled from the short-term impact of papers across four different citation networks. Our findings demonstrate that our method outperforms existing methods in terms of both metrics. Moreover, they validate the introduction of the attention-based mechanism. The effectiveness of our approach degrades when the attention-based mechanism is completely removed, or when used in isolation.

References

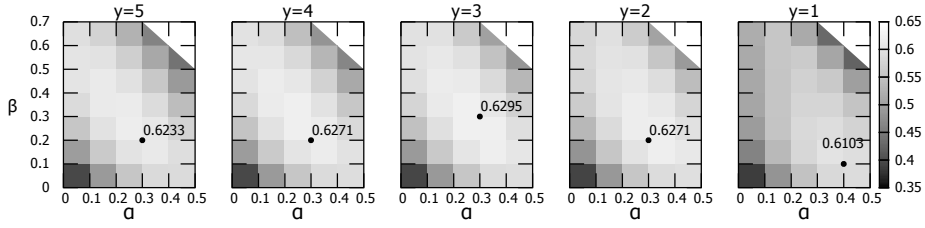
1. Bai, X., Liu, H., Zhang, F., Ning, Z., Kong, X., Lee, I., Xia, F.: An overview on evaluating and predicting scholarly article impact. Information **8**(3), 73 (2017). DOI 10.3390/info8030073
2. Barabási, A.L.: Network science book. Boston, MA: Center for Complex Network, Northeastern University. Available online at: http://barabasi.com/networksciencebook (2014)
3. Bergstrom, C.T., West, J.D., Wiseman, M.A.: The eigenfactor metrics. The Journal of Neuroscience **28**(45), 11433–11434 (2008)
4. Bornmann, L., Mutz, R.: Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references. Journal of the Association for Information Science and Technology **66**(11), 2215–2222 (2015)
5. Chang, S., Go, S., Wu, Y., Lee, Y., Lai, C., Yu, S., Chen, C., Chen, H., Tsai, M., Yeh, M., Lin, S.: An ensemble of ranking strategies for static rank prediction in a large heterogeneous graph. WSDM Cup (2016)
6. Chen, P., Xie, H., Maslov, S., Redner, S.: Finding scientific gems with google's pagerank algorithm. Journal of Informetrics **1**(1), 8–15 (2007)
7. Chin-Chi, H., Kuan-Hou, C., Ming-Han, F., Yueh-Hua, W., Huan-Yuan, C., Sz-Han, Y., Chun-Wei, C., Ming-Feng, T., Mi-Yen, Y., Shou-De, L.: Time-aware weighted pagerank for paper ranking in academic graphs. WSDM Cup (2016)
8. Davletov, F., Aydin, A.S., Cakmak, A.: High impact academic paper prediction using temporal and topological features. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM 2014, Shanghai, China, November 3-7, 2014, pp. 491–498. ACM (2014). DOI 10.1145/2661829.2662066
9. Diodato, V.P., Gellatly, P.: Dictionary of Bibliometrics (Haworth Library and Information Science). Routledge (1994)
10. Dunaiski, M., Visser, W.: Comparing paper ranking algorithms. In: Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, pp. 21–30. ACM (2012)
11. Feng, M., Chan, K., Chen, H., Tsai, M., Yeh, M., Lin, S.: An efficient solution to reinforce paper ranking using author/venue/citation information-the winner's solution for wsdm cup 2016. WSDM Cup (2016)
12. Garfield, E.: The history and meaning of the journal impact factor. Jama **295**(1), 90–93 (2006)
13. Ghosh, R., Kuo, T.T., Hsu, C.N., Lin, S.D., Lerman, K.: Time-aware ranking in dynamic citation networks. In: Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, pp. 373–380. IEEE (2011)
14. Hwang, W.S., Chae, S.M., Kim, S.W., Woo, G.: Yet another paper ranking algorithm advocating recent publications. In: Proceedings of the 19th international conference on World wide web, pp. 1117–1118. ACM (2010)
15. Jiang, X., Sun, X., Zhuge, H.: 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, pp. 714–723. ACM (2012)
16. Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T., Vassiliou, Y.: Impact-based ranking of scientific publications: A survey and experimental evaluation. Transactions on Knowledge and Data Engineering (to appear). DOI 10.1109/TKDE.2019.2941206

17. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. Journal of the ACM (JACM) **46**(5), 604–632 (1999)
18. Langville, A.N., Meyer, C.D.: Google's PageRank and beyond: The science of search engine rankings. Princeton University Press (2011)
19. Larsen, P., Von Ins, M.: The rate of growth in scientific publication and the decline in coverage provided by science citation index. Scientometrics **84**(3), 575–603 (2010)
20. Liao, H., Mariani, M.S., Medo, M., Zhang, Y.C., Zhou, M.Y.: Ranking in evolving complex networks. Physics Reports **689**, 1–54 (2017)
21. Ma, N., Guan, J., Zhao, Y.: Bringing pagerank to the citation analysis. Information Processing & Management **44**(2), 800–810 (2008)
22. Ma, S., Gong, C., Hu, R., Luo, D., Hu, C., Huai, J.: Query independent scholarly article ranking. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 953–964. IEEE (2018)
23. Nie, Z., Zhang, Y., Wen, J.R., Ma, W.Y.: Object-level ranking: bringing order to web objects. In: Proceedings of the 14th international conference on World Wide Web, pp. 567–574. ACM (2005)
24. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. rep., Stanford InfoLab (1999)
25. Sayyadi, H., Getoor, L.: Futurerank: Ranking scientific articles by predicting their future pagerank. In: Proceedings of the 2009 SIAM International Conference on Data Mining, pp. 533–544. SIAM (2009)
26. Shen, H., Wang, D., Song, C., Barabási, A.L.: Modeling and predicting popularity dynamics via reinforced poisson processes. In: Twenty-eighth AAAI conference on artificial intelligence (2014)
27. Sidiropoulos, A., Manolopoulos, Y.: Generalized comparison of graph-based ranking algorithms for publications and authors. Journal of Systems and Software **79**(12), 1679–1700 (2006)
28. Spearman, C.: The proof and measurement of association between two things. The American journal of psychology **15**(1), 72–101 (1904)
29. Wade, A.D., Wang, K., Sun, Y., Gulli, A.: Wsdm cup 2016: Entity ranking challenge. In: Proceedings of the ninth ACM international conference on web search and data mining, pp. 593–594. ACM (2016)
30. Walker, D., Xie, H., Yan, K.K., Maslov, S.: Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment **2007**(06), P06010 (2007)
31. Wang, D., Song, C., Barabási, A.L.: Quantifying long-term scientific impact. Science **342**(6154), 127–132 (2013)
32. Wang, J., Mei, Y., Hicks, D.: Comment on quantifying long-term scientific impact. Science **345**(6193), 149–149 (2014)
33. Wang, Y., Tong, Y., Zeng, M.: Ranking scientific articles by exploiting citations, authors, journals, and time information. In: AAAI (2013)
34. Wesley-Smith, I., Bergstrom, C.T., West, J.D.: Static ranking of scholarly papers using article-level eigenfactor (alef). arXiv preprint arXiv:1606.08534 (2016)
35. Xiao, S., Yan, J., Li, C., Jin, B., Wang, X., Yang, X., Chu, S.M., Zha, H.: On modeling and predicting individual paper citation count over time. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016, pp. 2676–2682. IJCAI/AAAI Press (2016)
36. Yan, E., Ding, Y.: Weighted citation: An indicator of an article's prestige. Journal of the Association for Information Science and Technology **61**(8), 1635–1643 (2010)
37. Yan, R., Tang, J., Liu, X., Shan, D., Li, X.: Citation count prediction: learning to estimate future citations for literature. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1247–1252. ACM (2011)
38. Yao, L., Wei, T., Zeng, A., Fan, Y., Di, Z.: Ranking scientific publications: the effect of nonlinearity. Scientific reports **4** (2014)
39. Yu, P.S., Li, X., Liu, B.: 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, pp. 543–549. IEEE (2005)
40. Zhang, F., Wu, S.: 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, pp. 127–130. ACM (2018)
41. Zhou, J., Zeng, A., Fan, Y., Di, Z.: Ranking scientific publications with similarity-preferential mechanism. Scientometrics **106**(2), 805–816 (2016)

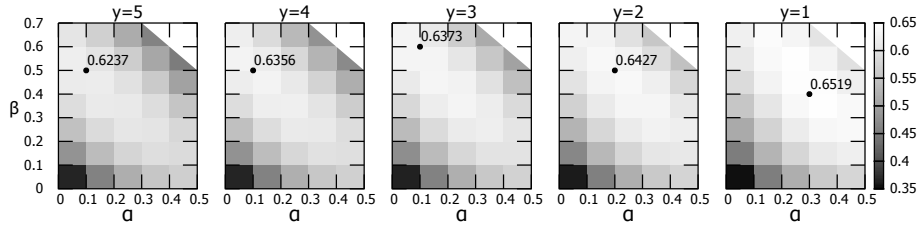
A Addendum To Evaluation

A.1 Ranking Effectiveness

Figure 6 presents heatmaps of AttRank's achieved correlation w.r.t. the ground truth STI ranking, on the APS and hep-th datasets using the default test ratio (see Section 4). We observe that in all cases, as β approaches 0 the correlations achieved drop dramatically (notice the darker hues on the bottom left corner of each figure), while the best correlation is always achieved for $\beta \neq 0$, illustrating the importance of the newly introduced attention-based vector.



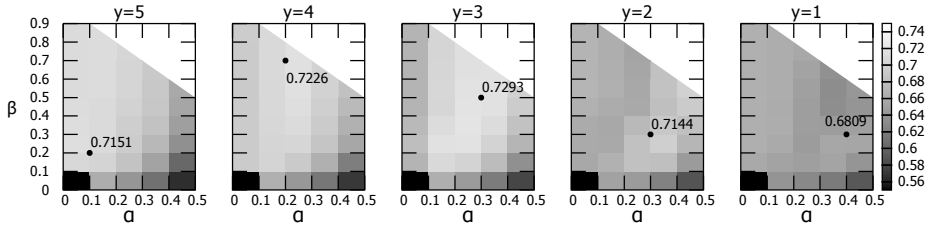
(a) Correlation on the APS dataset.



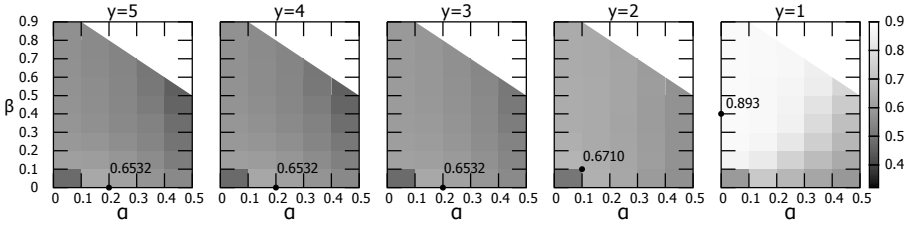
(b) Correlation on the hep-th dataset.

Fig. 6: Heatmaps depicting the effect of the parameterization of AttRank to its effectiveness in terms of the correlation for the APS and hep-th datasets. The best value achieved for each metric is also depicted.

Figure 7 presents heatmaps of AttRank's achieved nDCG@50 w.r.t. the ground truth STI ranking, on the APS and hep-th datasets using the default test ratio. The observations are as in the case of correlation: nCDG drops as β approaches 0, or as α increases towards 0.5, and its best value is achieved for some $\beta \neq 0$.



(a) nDCG@50 on the APS dataset.



(b) nDCG@50 on the hep-th dataset.

Fig. 7: Heatmaps depicting the effect of the parameterization of AttRank to its effectiveness in terms of nDCG@50 for the APS and hep-th datasets. The best value achieved for each metric is also depicted.