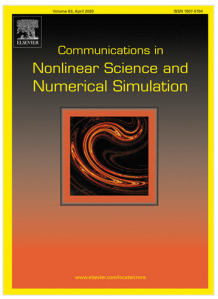
Algorithmic bias amplification via temporal effects: The case of PageRank in evolving networks

Cui, Mengtian ; Mariani, Manuel ; Medo, Matúš

Abstract: Biases impair the effectiveness of algorithms. For example, the age bias of the widely-used PageRank algorithm impairs its ability to effectively rank nodes in growing networks. PageRank's temporal bias cannot be fully explained by existing analytic results that predict a linear relation between the expected PageRank score and the indegree of a given node. We show that in evolving networks, under a mean-field approximation, the expected PageRank score of a node can be expressed as the product of the node's indegree and a previously-neglected age factor which can "amplify" the indegree's age bias. We use two well-known empirical networks to show that our analytic results explain the observed PageRank's age bias and, when there is an age bias amplification, they enable estimates of the node PageRank score that are more accurate than estimates based solely on local structural information. Accuracy gains are larger in degree-degree correlated networks, as revealed by a growing directed network model with tunable assortativity. Our approach can be used to analytically study other kinds of ranking bias.

DOI: https://doi.org/10.1016/j.cnsns.2021.106029

Journal Pre-proof



Algorithmic bias amplification via temporal effects: The case of PageRank in evolving networks

Mengtian Cui, Manuel Sebastian Mariani, Matúš Medo

PII: S1007-5704(21)00341-5
DOI: https://doi.org/10.1016/j.cnsns.2021.106029
Reference: CNSNS 106029

To appear in: *Communications in Nonlinear Science and Numerical Simulation*

Received date : 20 February 2021
Revised date : 22 July 2021
Accepted date : 31 August 2021

Please cite this article as: M. Cui, M.S. Mariani and M. Medo, Algorithmic bias amplification via temporal effects: The case of PageRank in evolving networks. *Communications in Nonlinear Science and Numerical Simulation* (2021), doi: https://doi.org/10.1016/j.cnsns.2021.106029.

Algorithmic bias amplification via temporal effects: The case of PageRank in evolving networks

Mengtian Cui[a], Manuel Sebastian Mariani[b], Matúš Medo[c,d]

[a]*Key laboratory of Computer System, State Ethnic Affairs Commission, Southwest Minzu University, 610041 Chengdu, P.R.China*
[b]*URPP Social Networks, Universität Zürich, Switzerland*
[c]*Department of Radiation Oncology, Inselspital, University Hospital of Bern, and University of Bern, 3010 Bern, Switzerland*
[d]*Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland*

Abstract

Biases impair the effectiveness of algorithms. For example, the age bias of the widely-used PageRank algorithm impairs its ability to effectively rank nodes in growing networks. PageRank's temporal bias cannot be fully explained by existing analytic results that predict a linear relation between the expected PageRank score and the indegree of a given node. We show that in evolving networks, under a mean-field approximation, the expected PageRank score of a node can be expressed as the product of the node's indegree and a previously-neglected age factor which can "amplify" the indegree's age bias. We use two well-known empirical networks to show that our analytic results explain the observed PageRank's age bias and, when there is an age bias amplification, they enable estimates of the node PageRank score that are more accurate than estimates based solely on local structural information. Accuracy gains are larger in degree-degree correlated networks, as revealed by a growing directed network model with tunable assortativity. Our approach can be used to analytically study other kinds of ranking bias.

Keywords: PageRank, Algorithmic bias, Networks, Ranking algorithms

1. Introduction

Algorithmic biases threaten the application of algorithms to real-world problems. Their presence can indeed penalize underrepresented minorities [1], reduce the success rate of higher-quality products in cultural markets [2], increase polarization and fragmentation in society [3], and, overall, lead to sub-optimal decisions. The major challenge toward understanding the biases of complex algorithms is that it might be hard to dissect the steps used by the algorithm to reach a certain decision [4]. But in complex environments, such as modern social and information systems, even the behavior of simpler algorithms might be difficult to understand [5], which makes it challenging to understand how algorithmic biases emerge and how to best prevent them.

A paradigmatic example of an algorithm whose biases are not yet fully understood is the widely-used PageRank algorithm [6, 7, 8]. The algorithm has been used for many and diverse applications, e.g., identifying seminal papers [9] and promising researchers [10] in academia, quantifying individual influence on collective opinion formation processes [6], predicting road traffic [11], and others [12]. The algorithm builds on the plausible hypothesis that in a given directed complex network (where the network's nodes are connected by directed edges), an important node is referred to by other important nodes [13]. Following the seminal paper that introduced it [7], PageRank has become a standard tool in network analysis, and it is often believed to provide more reliable and robust estimations of node importance than the number of incoming connections received by a node (referred to as indegree in network science) [8, 13].

Yet, numerical simulations and analyses of real-world information networks reveal that PageRank can be severely biased by the age of the nodes [14, 15, 16], and other factors (such as homophily effects [17] and domain [18]), which

Email address: manuel.mariani@business.uzh.ch (Manuel Sebastian Mariani)

impairs the algorithm's ability to identify high-potential nodes [15, 19, 20, 21, 22]. In networks with strong temporal effects [23], the age bias of the PageRank scores is more severe than that exhibited by the indegree [15, 19, 16]. At the same time, existing analytic results cannot explain why PageRank has a higher level of bias than indegree, as they predict that the PageRank score of a node is, on average, proportional to its indegree [24, 25]. Achieving an analytic understanding of the emergence of PageRank's biases remains elusive, but it might help to design robust methods to correct them.

Here, we provide an analytic explanation of PageRank biases in evolving networks. To this end, by performing a mean-field approximation, we derive a self-consistent equation for the PageRank score of nodes of a given indegree and age. We use this equation to show that in evolving networks, on average, the PageRank score of a node is expected to be proportional not only to its indegree, but also to a time-dependent factor, which we refer to as the *age factor*. The age factor depends on the temporal structure of the network. We use an empirical citation network from the Americal Physical Society to show that the age factor can "amplify" the indegree's age bias to produce PageRank scores that are considerably more biased than indegree. When an amplification effect is present, the obtained factorization of the expected PageRank score in a product of indegree and the age factor: (1) leads to an "age-corrected predictor" that estimates the PageRank score of a node more accurately than indegree alone and (2) provides support for the practice to "debias" PageRank scores using rescaling procedures [19, 20]. In an empirical network without amplification (i.e., where the age factor is close to one), represented here by the Higgs Twitter dataset [26], the indegree alone provides an accurate estimate of the PageRank score. Indegree assortativity plays a key role in the relative accuracy of the age-corrected and the indegree predictors, as revealed by systematic analysis of synthetic data generated with a growing network model.

2. Background

This Section provides background information that motivates the rest of the paper. It introduces the definitions of indegree and PageRank, previous analytic arguments [24, 25] pointing at a linear relation between the expected PageRank score of a node and its indegree, and previous numeric evidence that suggests instead a more complex relation between PageRank and indegree [15, 19].

2.1. Indegree and PageRank

Given a directed network composed of N nodes whose $N \times N$ adjacency matrix is denoted by A ($A_{ij} = 1$ if a directed edge from j to i exists, zero otherwise), the indegree $k_{in}(i)$ of a given node i is defined as the total number of its incoming edges, $k_{in}(i) = \sum_j A_{ij}$ [16]. When there are no dangling nodes (i.e., no nodes with outdegree equal to zero), the vector $\boldsymbol{x}$ of PageRank scores is defined as the stationary state of the following process [12]:

$$\boldsymbol{x}_{n+1} = \alpha\, \mathsf{P}\, \boldsymbol{x}_n + (1-\alpha)\, \boldsymbol{v}, \tag{1}$$

where P denotes the transition matrix whose elements are $P_{ij} = A_{ij}/k_{out}(j)$ where $k_{out}(j) = \sum_i A_{ij}$ denotes node j's outdegree, and $\boldsymbol{v}$ is a teleportation vector which, in line with many applications of the algorithm [12], we assume to be uniform, $v(i) = 1/N$. Component by component, Eq. (1) reads[1]

$$x_{n+1}(i) = \alpha \sum_j \frac{A_{ij}}{k^{out}(j)}\, x_n(j) + \frac{1-\alpha}{N}. \tag{2}$$

Intuitively, the equation embodies the idea that an important node is referred to by other important nodes, as i's score depends linearly on its neighbors' scores. The nodes' final scores, $\boldsymbol{x} = (x(1), \ldots, x(N))$, can be obtained by the power method [16], i.e., by starting from an initial condition and recursively applying Eq. (2) until the L_1 distance between $\boldsymbol{x}_{n-1}$ and $\boldsymbol{x}_n$ becomes smaller than a predefined tolerance parameter, ϵ [8]. The convergence conditions for Eq. (2)

[1] Incorporating dangling nodes would simply require, in Eq. (2), to replace the first term of the r.h.s. with the sum of a contribution from the non-dangling nodes ($\sum_{i:k_{out}>0} A_{ij}\, x_n(j)/k^{out}(j)$) and a uniform contribution from all the dangling nodes – see [8] for a detailed discussion. The contribution from the dangling nodes would affect all nodes in the same way and, for this reason, it can be omitted from the analysis with no loss of generality [24].

have been well-studied in the literature [8]. In fact, Eq. (2) can be interpreted as the master equation of a Markov chain with transition probability matrix $\mathsf{G} = \alpha\mathsf{P} + (1-\alpha)\boldsymbol{e}^T\boldsymbol{e}/N$, where P is the matrix whose element is defined as $P_{ij} = A_{ij}/k^{out}(j)$. As matrix G (typically referred to as Google matrix) is both stochastic and primitive, a unique solution of Eq. (2) exists, and iterating Eq. (2) will reach it independently of the initial condition [8].

2.2. Relation between PageRank and indegree

By simply looking at Eq. (2), one could conjecture that nodes with many incoming edges have a larger PageRank score. Fortunato et al. [24] provided analytic results to support this intuition using a mean-field approximation. Specifically, they estimated the average PageRank score, $\overline{x(\boldsymbol{k})}$, of a given set of nodes (called *degree class*) characterized by a given pair of indegree and outdegree values. We denote a degree class as $C(\boldsymbol{k}) = C(k_{in}, k_{out}) = \{i : k_{in}(i) = k_{in} \text{ and } k_{out}(i) = k_{out}\}$. With this notation, the average PageRank score of nodes in degree class $\mathbf{k}$ is defined as:

$$\overline{x(\boldsymbol{k})} = \frac{1}{N\,P(\boldsymbol{k})} \sum_{i \in C(\boldsymbol{k})} x(i), \tag{3}$$

where $N\,P(\boldsymbol{k})$ is the total number of nodes of degree class $C(\boldsymbol{k})$, and $P(\boldsymbol{k}) = P(k_{in}, k_{out})$ denotes the joint degree distribution of the network (defined as the fraction of nodes with indegree and outdegree equal to k_{in} and k_{out}, respectively) [27]. Under a mean field approximation, one assumes that the PageRank score of a neighbor of class $C(\boldsymbol{k}')$ can be replaced by the average PageRank of nodes of that class. Therefore,

$$\sum_{i \in C(\boldsymbol{k})} \sum_{j \in C(\boldsymbol{k}')} A_{ij}\, x_n(j) \simeq \overline{x_n(\boldsymbol{k}')}\, E_{\boldsymbol{k}' \to \boldsymbol{k}}, \tag{4}$$

where $E_{\boldsymbol{k}' \to \boldsymbol{k}}$ is the number of edges from nodes of class $\boldsymbol{k}'$ to nodes of class $\boldsymbol{k}$. Using this approximation, from Eq. (2), one can obtain the following self-consistent equation for the average PageRank score:

$$\overline{x_{n+1}(\boldsymbol{k})} = \alpha\, k_{in} \sum_{C(k')} \frac{P_{in}(k'|\boldsymbol{k})}{k'_{out}}\, \overline{x_n(\boldsymbol{k}')} + \frac{1-\alpha}{N}, \tag{5}$$

where $P_{in}(\boldsymbol{k}'|\boldsymbol{k})$ is the probability that a link to a node in class $C(k)$ originated from a node in class $C(\boldsymbol{k}')$. For networks without degree correlations among neighboring nodes (which we refer to as *uncorrelated networks* from now on), one has $P_{in}(\boldsymbol{k}'|\boldsymbol{k}) = P(\boldsymbol{k}')\,k'_{out}/\langle k_{out}\rangle$, which leads to the following proportionality relation between average PageRank score and indegree [24]:

$$\overline{x(\boldsymbol{k})} = \frac{\alpha}{N}\frac{k_{in}}{\langle k_{in}\rangle} + \frac{1-\alpha}{N}. \tag{6}$$

Eq. (6) is consistent with our expectation that if a node has many incoming edges, it is likely to achieve a high PageRank score as a result. However, this result neglects temporal patterns in network connectivity, which may strongly affect the relation between PageRank and indegree. Understanding the role of temporal patterns for the analytic relation between the PageRank score and indegree is our main goal.

2.3. PageRank's biases and limitations of existing works

Although the analytic argument revisited above suggests that the PageRank score might behave similarly to indegree, numeric results suggest that the relation between PageRank and indegree is more complex than the simple linear relation expected from Eq. (6).

In citation networks where each node can only refer to older nodes (e.g., a paper citing an existing paper), PageRank's bias toward old nodes is more severe than indegree's bias. Considering the citation network of the papers published in Americal Physical Society (APS), for example, a previous work [19] found that the proportion of old (recent) nodes at the top of the ranking by PageRank is substantially larger (smaller) than the same proportion at the top of the ranking by indegree. A similar finding holds for citation networks of patents [20]. Yet, these works only reported the relative magnitude of PageRank's and indegree's temporal biases in empirical citation data, without clarifying the reasons behind the different magnitude of the two metrics' biases.

Refs. [19, 20] and some follow-ups [18, 28, 21] showed that we can effectively suppress both indegree's and PageRank's temporal biases through a simple rescaling procedure. In the proposed rescaling procedure, a node's score is only compared against the scores of nodes of similar age using a z-score calculation [19]. Although the rescaling procedure is effective at suppressing PageRank's temporal bias [19, 20], previous works did not attempt to justify it analytically.

To uncover the mechanisms behind the emergence of PageRank's bias, Ref. [15] adopted a simulation-based approach. In simulations where the nodes of a growing network create outgoing links over an extended time period, they found that PageRank can be biased toward either recent or old nodes depending on the relatives timescale at which the creation of outgoing links and the attraction of incoming links decay [15]. However, this study only reported the observed relation between PageRank and indegree in numeric simulations, together with their biases, without attempting to analytically explain the observed relation.

In this paper, we overcome the limitations of previous works by proposing an analytic approach to quantifying the relation between PageRank and indegree in evolving networks. The proposed approach identifies the circumstances under which PageRank is expected to be more biased by node age than indegree, and it clarifies why a simple rescaling procedure of the nodes' PageRank scores is effective at suppressing their age bias.

3. Analytic results

In this Section, we focus on the relation between PageRank and indegree in evolving networks. We derive a general self-consistent equation that links the expected PageRank score, indegree, and the temporal structure of the network data. We use a mean-field approximation to show that the expected PageRank score of a node can be decomposed as the product of its indegree and an age factor that depends on the temporal structure of the data. We finally show how rescaling PageRank scores [19] can factor out their age dependence.

3.1. Linking PageRank score, indegree, and temporal effects

This article focuses on *evolving networks*, broadly defined as networks that change in time through the addition of new nodes and edges and the deletion of preexisting ones. Many real networks are evolving ones, including citation networks of scientific papers and patents, online and offline social networks where new ties are incessantly created and deleted, and user-product networks in e-commerce systems, among others [23]. In all these networks, the propensity that two nodes are connected depends on their age. For example, in a citation network, a link from a node i to j can only exist if i is younger than j. To formalize this property, we divide the N nodes into $B = N/\Delta$ age classes composed of Δ nodes each, denoted as $C(t)$ ($t = 1, \ldots, B$). As in Section 2.2, we assume that a node belongs to degree class $C(\boldsymbol{k}) = C(k^{in}, k^{out})$ if its indegree and outdegree are equal to k^{in} and k^{out}, respectively. Further, we split the nodes into *degree-age classes*: Degree-age class $C(\boldsymbol{k}, t) = C(\boldsymbol{k}) \cap C(t)$ comprises the nodes of degree class $C(\boldsymbol{k})$ and age class $C(t)$. For a generic evolving network, we assume that the temporal structure of the data is encoded in a $B \times B$ *temporal transition matrix* whose element $P(t' \to t)$ denotes the probability that an edge from a node in class $C(t')$ links to a node in age class $C(t)$.

We assume that a snapshot of the evolving network is given, and we set out to estimate the average PageRank score $\overline{x(\boldsymbol{k}, t)}$ of a node of degree-age class $C(\boldsymbol{k}, t)$ from the sole properties of class $C(\boldsymbol{k}, t)$. Specifically, we define

$$\overline{x(\boldsymbol{k}, t)} := \frac{1}{N\, P(\boldsymbol{k}, t)} \sum_{i \in C(\boldsymbol{k}, t)} x(i), \tag{7}$$

where $P(\boldsymbol{k}, t)$ is the fraction of nodes that belong to class $C(\boldsymbol{k}, t)$. Subsequently, we plug Eq. (2) into this definition, and we use the property that summing over all nodes is equivalent to first summing over all degree-age classes, and then summing over all nodes that belong to each degree-age class. We obtain

$$\overline{x_{n+1}(\boldsymbol{k}, t)} = \frac{1}{N\, P(\boldsymbol{k}, t)} \sum_{i \in C(\boldsymbol{k}, t)} \left(\alpha \sum_j \frac{A_{ij}}{k^{out}(j)}\, x_n(j) + \frac{1-\alpha}{N} \right) = \frac{\alpha}{N\, P(\boldsymbol{k}, t)} \sum_{C(\boldsymbol{k}', t')} \sum_{i \in C(\boldsymbol{k}, t)} \sum_{j \in C(\boldsymbol{k}', t')} A_{ij}\, \frac{x_n(j)}{k_{out}(j)} + \frac{1-\alpha}{N}. \tag{8}$$

At this stage, similarly to Eq. (4), we perform a mean-field approximation where we assume that the PageRank score of a given neighbor j can be replaced by the average PageRank score of the class to which j belongs. Hence

$$\sum_{i\in C(\boldsymbol{k},t)} \sum_{j\in C(\boldsymbol{k}',t')} A_{ij}\, x_n(j) \simeq \overline{x_{n+1}(\boldsymbol{k}',t')}\, E_{(\boldsymbol{k}',t')\to(\boldsymbol{k},t)} \tag{9}$$

where $E_{(\boldsymbol{k}',t')\to(\boldsymbol{k},t)}$ denotes the total number of edges from nodes in class $C(\boldsymbol{k}',t')$ to nodes in class $C(\boldsymbol{k},t)$. This quantity can be expressed as

$$E_{(\boldsymbol{k}',t')\to(\boldsymbol{k},t)} = N\, k_{in}\, P(\boldsymbol{k},t)\, P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t), \tag{10}$$

where $P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t)$ denotes the probability that a directed edge to a node in $C(\boldsymbol{k},t)$ comes from a node in $C(\boldsymbol{k}',t')$. In the following, we will also denote as $P_{in}(\boldsymbol{k}',t'\to\boldsymbol{k},t)$ the probability that a directed edge from a node in $C(\boldsymbol{k}',t')$ links to a node in $C(\boldsymbol{k}',t')$. Note that $P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t)$ and $P_{in}(\boldsymbol{k}',t'\to\boldsymbol{k},t)$ differ. Indeed, by denoting as $E_{(\boldsymbol{k}',t')\to\cdot}$ and $E_{\cdot\to(\boldsymbol{k},t)}$ the total number of edges from nodes in $C(\boldsymbol{k}',t')$ and the total number of edges to nodes in $C(\boldsymbol{k},t)$, respectively, we have $P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t) = E_{(\boldsymbol{k}',t')\to(\boldsymbol{k},t)}/E_{\cdot\to(\boldsymbol{k},t)}$ and $P_{in}(\boldsymbol{k}',t'\to\boldsymbol{k},t) = E_{(\boldsymbol{k}',t')\to(\boldsymbol{k},t)}/E_{(\boldsymbol{k}',t')\to\cdot}$. Plugging Eqs. (9) and (10) into Eq. (8), we obtain

$$\overline{x_{n+1}(\boldsymbol{k},t)} = \alpha\, k_{in} \sum_{C(\boldsymbol{k}',t')} \frac{P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t)}{k'_{out}}\, \overline{x_n(\boldsymbol{k}',t')} + \frac{1-\alpha}{N}. \tag{11}$$

This is a self-consistent equation for the average PageRank score of nodes of degree-age class $C(\boldsymbol{k},t)$. To further simplify it, we will focus on degree-uncorrelated networks.

3.2. PageRank, indegree, and age factor in degree-uncorrelated networks

We focus on degree-uncorrelated evolving networks. Specifically, we assume that the probability that an edge to a node in $C(\boldsymbol{k},t)$ comes from a node in $C(\boldsymbol{k}',t')$ only depends on the probability $P_{in}(t'|t)$ that an edge to a node in age class $C(t)$ comes from a node in age class $C(t')$, times the fraction of outgoing edges in $C(t')$ from nodes with outdegree k'_{out}; the latter factor enforces the assumption of a lack of degree-degree correlations. By further assuming that node outdegree is independent of node age, we have

$$P_{in}(\boldsymbol{k}',t'|\boldsymbol{k},t) = P_{in}(t'|t)\, \frac{P(k'_{out})\, k'_{out}}{\langle k_{out}\rangle}, \tag{12}$$

where $P(k'_{out})\, k'_{out}/\langle k_{out}\rangle$ denotes the fraction of outgoing edges that come from nodes of outdegree k'_{out}. Plugging Eq. (12) into Eq. (11), we obtain

$$\overline{x_{n+1}(\boldsymbol{k},t)} = \alpha\, k_{in} \sum_{(\boldsymbol{k}',t')} \frac{P_{in}(t'|t)\, P(k'_{out})}{\langle k_{out}\rangle}\, \overline{x_n(\boldsymbol{k}',t')} + \frac{1-\alpha}{N}. \tag{13}$$

To obtain a simple analytic dependence of $\overline{x_{n+1}(\boldsymbol{k},t)}$ on indegree, we assume that all nodes have the same outdegree m, i.e., $P(k_{out}) = \delta_{k_{out},m}$ where δ denotes the Kronecker delta. This assumption is justified in systems where the outdegree distribution is substantially narrower than the indegree distribution. The networks analyzed here respect this property (see Section 4.2). We shall mention at the end of this Section the main difference in the final result if we were to take into account the outdegree heterogeneity.

Under the homogeneous-outdegree assumption, the degree-age classes can be denoted as $C(k_{in},t)$, and the outdegree is omitted from the notation. With this assumption, the sum in Eq. (13) can be rewritten as

$$\sum_{C(k'_{in},t')} P_{in}(t'|t)\, \overline{x_n(k'_{in},t')} \simeq \sum_{C(t')} P_{in}(t'|t) \sum_{C(k'_{in})} P(k'_{in},t')\, \overline{x_n(k'_{in},t')} = \sum_{C(t')} P_{in}(t'|t)\, \overline{x_n(t')} = \sum_{t'} P_{in}(t'|t)\, \overline{x_n(t')}, \tag{14}$$

where $\overline{x_n(t')}$ is the average PageRank score of nodes in age class $C(t')$. By plugging Eq. (14) into Eq. (13), we obtain

$$\overline{x_{n+1}(\boldsymbol{k},t)} = \frac{\alpha}{m}\, k_{in} \sum_{t'} P_{in}(t'|t)\, \overline{x_n(t')} + \frac{1-\alpha}{N}. \tag{15}$$

For a more straightforward interpretation, we express $P_{in}(t'|t)$ in terms of the probability $P(t' \to t)$ that a node in $C(t')$ links to a node in $C(t)$. We have:

$$P_{in}(t'|t) = P(t' \to t)\,\frac{m}{\langle k_{in}(t)\rangle} \tag{16}$$

where $\langle k_{in}(t)\rangle$ is the average indegree of nodes in age class $C(t)$. This leads to the expression for the average PageRank score of nodes in $C(k_{in}, t)$

$$\overline{x_{n+1}(k_{in}, t)} = \alpha\,\frac{k_{in}}{\langle k_{in}(t)\rangle}\sum_{t'} P(t' \to t)\,\overline{x_n(t')} + \frac{1-\alpha}{N}, \tag{17}$$

where $\overline{x_n(t')}$ can be found by solving the iterative equation

$$\overline{x_{n+1}(t)} = \frac{\alpha}{\Delta}\sum_{t'} P(t' \to t)\,\overline{x_n(t')} + \frac{1-\alpha}{N}. \tag{18}$$

Hence, Eq. (17) can be conveniently rewritten as:

$$\overline{x(k_{in}, t)} = \alpha\; k_{in}\,\phi_{in}(t) + \frac{1-\alpha}{N}, \tag{19}$$

where the age factor

$$\phi_{in}(t) = \frac{1}{\langle k_{in}(t)\rangle}\sum_{t'} P(t' \to t)\,\overline{x_n(t')} \tag{20}$$

denotes the average "premium" that a node in class t obtains from the temporal structure of the data, encoded in the transition matrix $P(t' \to t)$. This is our first main result: In the first approximation, for uncorrelated networks where each node has the same outdegree, the average mean-field PageRank score of a node of indegree-age class $C(k_{in}, t)$ is proportional to the product of its indegree and an age factor that depends on the temporal structure of the network data.

We conclude this Section with two remarks. First, for networks where the formation of new edges does not depend on node age, the temporal transition matrix is uniform ($P(t' \to t) = 1/B$), which implies $\phi_{in}(t) = \langle k_{in}(t)\rangle^{-1} = \langle k_{in}\rangle^{-1}$ and, therefore, $\overline{x(k_{in}, t)} = \alpha\; k_{in}/\langle k_{in}\rangle + (1-\alpha)/N$, which is the analytic result found in prior works [24, 25]. Second, if we consider the outdegree heterogeneity, we would obtain an expression similar to Eq. (17) but without a simple interpretation. Specifically, in the sum on the r.h.s., we would not have the average PageRank of nodes of age class $C(t')$, $\overline{x_n(t')}$, but a more complicated expression depending on both the outdegree distribution and the joint distribution $P(\boldsymbol{k}, t)$.

3.3. Debiasing PageRank's scores by rescaling

Eq. (19) above has implications for the previously-studied PageRank's age bias [9, 14, 15, 19, 20]. It indicates indeed that PageRank's age bias is simultaneously affected by both the age bias of indegree, and the bias in the age factor, $\phi_{in}(t)$. In Section 3.1, we will illustrate this point with a concrete example. In the following, we show that the previous analytic results shed light on why a simple rescaling procedure is effective to mitigate PageRank's age bias. Recent works [19, 16, 29, 20, 21] defined the rescaled PageRank score of a node i, $R_x(i)$, as the number of standard deviations its PageRank score exceeds the mean PageRank score of nodes of similar age [19]:

$$R_x(i) = \frac{x - \overline{x(i)}}{\sigma[x(i)]}, \tag{21}$$

where $\overline{x(i)}$ and $\sigma[x(i)]$ denote the mean and standard deviation, respectively, of the PageRank scores of nodes of similar age as node i. Prior works determined this set of nodes as $[i - \delta/2, i + \delta/2]$, where δ is a parameter of the rescaling method [19]. For simplicity, in the following analytic developments, we assume that each node is compared against the nodes that belong to its age class. Therefore, we define the average rescaled PageRank score of a node in class $C(k_{in}, t)$, $\overline{R_x(k_{in}, t)}$, as

$$\overline{R_x(k_{in}, t)} = \frac{\overline{x(k_{in}, t)} - \overline{x(t)}}{\sigma[\overline{x(t)}]}. \tag{22}$$

For a node in class $C(k_{in}, t)$, from Eq. (19), we obtain

$$\begin{aligned} \overline{x(t)} &= (1-\alpha)[\alpha\ \langle k^{in}(t)\rangle\ \phi_{in}(t) + 1]/N, \\ \sigma[\overline{x(t)}] &= \alpha\, \sigma[k_{in}(t)]\, \phi_{in}(t), \end{aligned} \tag{23}$$

which implies that the average rescaled PageRank score of a node in class $C(k_{in}, t)$, $\overline{R_x(k_{in}, t)}$, is given by

$$\overline{R_x(k_{in}, t)} = \frac{k_{in} - \langle k_{in}(t)\rangle}{\sigma[k_{in}(t)]} = R_k(k^{in}, t). \tag{24}$$

This is our second main result: Rescaling the PageRank score eliminates the age factor, which leads the expected rescaled PageRank score of nodes in class $C(k_{in}, t)$ to be equal to their rescaled indegree score. Having the rescaling procedure factored out the age dependence that is specific to the PageRank score, if the rescaled indegree score is age-unbiased (which is usually the case [19, 20]), we expect the rescaled PageRank score to be unbiased too.

4. Results on empirical networks

We investigate the implications of the analytic results above in two empirical networks. We first introduce the procedures to compute the empirical PageRank scores and to estimate the expected PageRank scores using the analytic results above, and then investigate the agreement between the analytic estimates of expected PageRank scores and their empirical values.

4.1. Empirical datasets

Suitable empirical networks for our analysis need to be directed and time-stamped. We analyze two empirical datasets: The well-studied APS citation network [9, 14, 15, 19, 21] and a dataset on the Twitter activity related to the experimental discovery of the Higgs boson [26]. We use these datasets to illustrate how PageRank's age bias can be explained in terms of the analytic results above, and identify the circumstances under which the age-corrected predictor outperforms the indegree predictor at estimating a node's PageRank score.

APS dataset. We analyze a large-scale citation network of papers published in the American Physical Society (APS) journals (available under request at `https://journals.aps.org/datasets`), updated until year 2009. The networks includes 449, 935 papers and 4, 672, 812 citations between them. PageRank has been already applied to this dataset in prior studies [9, 14, 15, 19, 21], which revealed its age bias and the ability of rescaled PageRank to early identify editor-selected milestone papers [19] (see Section 4.1.2 in [29] for a review).

Higgs (Twitter) dataset. We analyze the dataset collected from Twitter by De Domenico et al. [26] to analyze the spreading of a scientific rumor – the experimental discovery of the Higgs boson, announced on 4th July 2012. The dataset is restricted to the period from 1st to 7th of July 2012, and it is publicly available at `https://snap.stanford.edu/data/higgs-twitter.html`. Here, we focus on the directed social network where a directed link from i to j means that i mentioned j. The resulting network has 116, 408 nodes and 145, 465 directed edges. Previous studies have used this dataset as a benchmark to study rumor spreading and influential users of online social networks [30].

4.2. Empirical and estimated PageRank scores

In both the APS and Higgs datasets, to compute the empirical PageRank values, we split the nodes into $B = 40$ age classes, and we compute the nodes' empirical scores using the power method adopted in prior works [15, 19] (i.e., by iterating Eq. (2)), with teleportation parameter $\alpha = 0.5$ and tolerance $\epsilon = 10^{-9}$, where $\alpha = 0.5$ is the customary choice for citation data [9, 19]. The analytic estimates of PageRank scores are obtained either through Eq. (6) (which we refer to as *indegree predictor*) or the novel Eq. (19) (*age-corrected predictor*). To estimate the age factor in Eq. (19), we measure the temporal transition matrix $P(t' \to t)$ from the empirical citation/mention flows between the 40 age classes (see Fig. 1), and then use it to find the expected PageRank of the nodes of a given age class through Eq. (18), which is then the input of Eq. (20) to determine the age factor $\phi_{in}(t)$. The age factor was derived

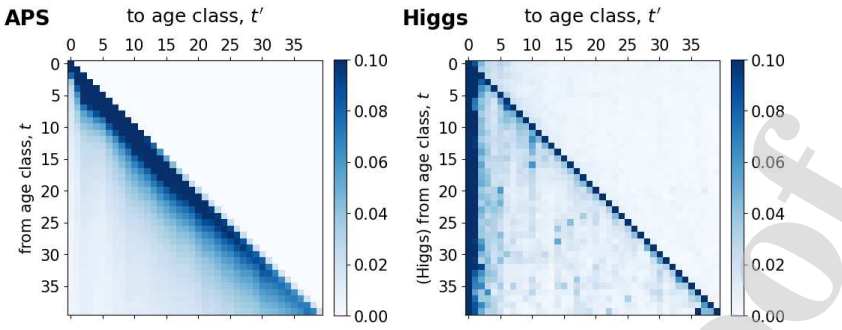


Figure 1: **Temporal transition matrix in the empirical networks.** For each pair of age classes t, t', the corresponding cell's color indicates the value of $P(t' \to t)$. This matrix is a key element for the analytic estimation of PageRank scores in evolving networks. In the APS network (left panel), the matrix $P(t' \to t)$ reveals two general properties of citation networks [29]: (1) nodes almost always cite nodes from previous age classes, (2) because of well-documented aging effects [15], nodes tend to cite nodes from near age classes. The shape of the matrix in the Higgs dataset (right panel) suggests that during the studied event, nodes tended to mention the earliest nodes that tweeted about the new particle's discovery, without substantial aging effects.

under the simplifying assumption of homogeneous outdegree, which is violated in most real networks. For the two empirical datasets analyzed here, a potential argument to support the approximation is that the outdegree distribution is substantially narrower than the indegree distribution, as evident by comparing the variance of the indegree against that of the outdegree ($\sigma^2(k_{out})/\sigma^2(k_{in})$ is 0.10 and $1.3 \cdot 10^{-3}$ in the APS and Higgs dataset, respectively).

4.3. Age bias amplification

Prior works found that in citation networks such as the APS network, PageRank's bias by age is more severe than indegree's bias, meaning that old papers are more overrepresented at the top of the ranking by PageRank than at the top of the ranking by indegree [15, 19]. The analytic result in Eq. (19) indicates that approximately, the PageRank score is determined by both the indegree of a node and an age factor. The precise shape of the age factor determines whether PageRank is more or less biased by age than indegree.

To quantify the age bias of a given metric across the age classes of papers, we compute the metric's average score for each age bin and normalize it with the metric's average score overall. Thus-obtained score fold increase can be used to assess the average age bias of the metric toward papers from a given age bin. A value greater (smaller) than one indicates that papers in a given bin are favored (disfavored), on average, by the evaluated metric. Note that other measures of ranking bias, such as those that focus on the best-ranked nodes [21], are possible.

In the APS network analyzed here, both the average indegree and the average PageRank score of a node are biased in favor of old nodes (Fig. 2A). Consistently with prior works, we find that PageRank's bias is more severe than indegree's bias, which is illustrated by the stronger advantage (disadvantage) of old nodes (recent nodes) compared to more recent (older) nodes. This can be explained by the fact that the age factor, ϕ_{in}, is heavily biased toward old nodes (see the orange line in Fig. 2A). By combining empirical age factor and average indegree values according to Eq. (19), we obtain an analytic prediction (dark blue line in Fig. 2A) of the temporal dependence of the average PageRank score, which matches accurately the empirical behavior (light blue line in Fig. 2A; the light blue shadowed area indicates the standard error of the mean of the score fold increase for the empirical PageRank score within each age class).

A different scenario is observed in the Higgs network. Here, the age factor is reasonably uniform across different age classes and it is close to one (Fig. 2C). This property indicates that although age classes corresponding to the oldest nodes accrue most of the links (see Fig. 1, right panel), their advantage is not further amplified by the structure of the temporal transition matrix in Eq. (20). According to Eq. (19), $\phi_{in}(t) \simeq 1$ predicts that the PageRank score will exhibit an age bias of similar magnitude than that exhibited by the degree, which is indeed confirmed by the empirical values of the PageRank score (Fig. 2C).

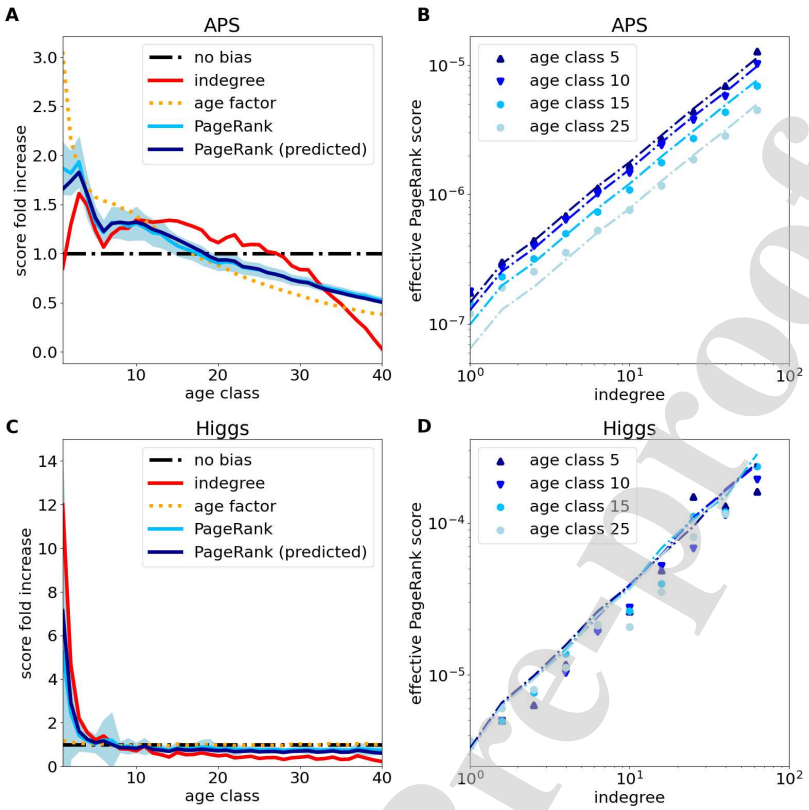


Figure 2: **Explaining PageRank's age bias in the empirical networks.** (A, C) Score fold increase of indegree, age factor $\phi_{in}(t)$, and PageRank for 40 age classes of papers, sorted from the oldest one ($t = 1$, left) to the most recent one ($t = 40$, right). In the APS network (panel A), compared to the indegree's age bias, the stronger bias exhibited by PageRank is a combination of the indegree's bias and the age factor. (B, D) The relation between the average PageRank and indegree for four different age classes. In the APS network (panel B), the relation is mediated by the age factor, and the mean-field analytic results (lines) match the empirical results (dots) well for sufficiently large indegree values. In the Higgs network (panel D), we observe no systematic discrepancies among the age classes, as predicted by the age factor.

4.4. The age factor moderates the relation between PageRank and indegree

An important prediction of Eq. (19) is that differently from mean-field results that do not consider the temporal structure of the data [24, 25], the relation between PageRank and indegree is moderated by the age factor, $\phi_{in}(t)$. To test this result, in both empirical datasets, we divide the nodes into 10 indegree classes (by adopting a logarithmic binning), and plot their average effective PageRank score, $\overline{x(k_{in}, t)} - (1 - \alpha)/N$, as a function of indegree for different age classes. According to the analytic result in Eq. (19), for a given age class t, we expect the relation between $\overline{x(k_{in}, t)}$ and k_{in} to be linear, with an age-dependent slope proportional to $\phi_{in}(t)$.

In the APS dataset, we find that for sufficiently large indegree values, the prediction in Eq. (19) accurately matches the empirical relation between PageRank score and indegree (Fig. 2B). The slope of the relation, manifested by the vertical displacement in the log-log scale of Fig. 2B, depends on the node age. The major discrepancies between the analytic and the empirical relationships are found for small indegree nodes, which is likely to be a consequence of the performed mean-field approximations (see [24] for similar discrepancies at small indegree values). Overall, compared to prior analytic results [24, 25], these findings confirm the analytic expectation that the slope of the linear relation between mean PageRank score and indegree depends on the node age.

In the Higgs dataset, the age factor is nearly uniform across different age classes, *i.e.*, $\phi_{in} \simeq 1$. In this case, our analytic computation predicts that the dependence of the PageRank score on indegree is approximately independent of the considered age class. We observe indeed no age-dependence of the relation between PageRank and indegree (Fig. 2D). This insight suggests that improving PageRank estimations by incorporating temporal information might

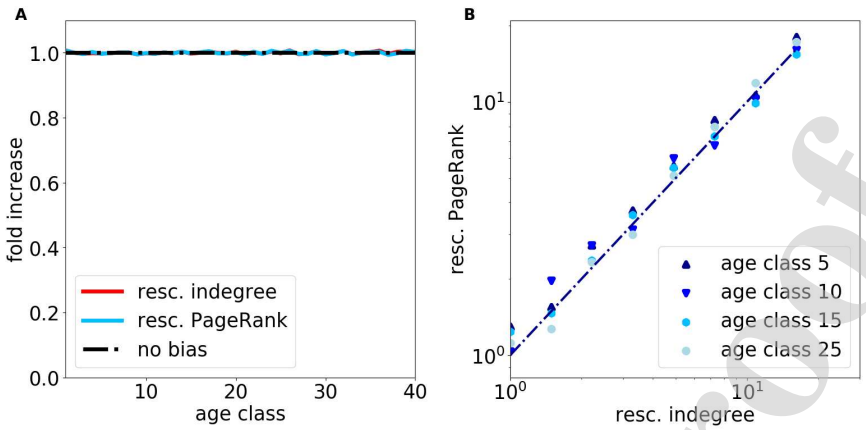


Figure 3: **Understanding the effects of the rescaling procedure.** (A) Score fold increase of rescaled indegree and rescaled PageRank for 40 age classes of papers. We see that both rescaled indegree and PageRank are essentially unbiased. (B) The relation between average rescaled PageRank and rescaled indegree for four different age classes. The four age classes follow the same curve that is well described by a direct proportionality relationship (shown as the dash-dotted line).

be challenging in the Higgs dataset.

4.5. Estimating PageRank from local and temporal information

After having analyzed the moderator role of age in the relation between PageRank and indegree, we focus on the problem of estimating the PageRank score of a given node by only using the properties of its degree-age class. We find that that the indegree predictor (Eq. 6) has good explanatory power: Its coefficient of determination of the empirical PageRank score in the APS data is $R^2 = 0.67$. This value is roughly consistent with previously-reported correlation values found in Web graphs [24, 25]. In the Higgs dataset, such correlation is even higher ($R^2 = 0.96$). The results in Figs. 2B and 2D suggest that by incorporating the age factor, we might achieve a substantially more accurate estimation of node-level PageRank score in the APS dataset, but not in the Higgs dataset.

We confirm that this is the case. In the APS dataset, the age-corrected predictor given by Eq. (19) has a stronger correlation with the empirical PageRank score ($R^2 = 0.79$, +18% compared to the indegree predictor). The fraction of variance unexplained by the combined age-indegree relation, $1 - R^2$, is due to the intrinsic structure of the citation network which, for example, involves degree correlations that the derivation of Eq. (19) neglects (see Sec. 4.7 for details). A qualitatively similar result can be obtained by measuring the mean squared errors of a given predictor, defined as $e(\overline{x}) = \sqrt{N^{-1}\sum_{i=1}^{N}(\overline{x}_i - x_i)^2}$, where x_i and $\overline{x}_i$ denote the empirical and predicted PageRank score, respectively. We find that compared to the indegree predictor ($e = 2.33 \times 10^{-6}$), the age-corrected predictor achieves a substantially lower error ($e = 1.96 \times 10^{-6}$, −16%).

By contrast, in the Higgs dataset, the age-corrected predictor has a similar correlation with the empirical PageRank score ($R^2 = 0.96$) to the indegree predictor. This suggests a limitation to the range of usefulness of our calculation: When the age factor is close to one for most age groups, the age-corrected predictor does not outperform the indegree predictor. A likely lack of improvement is signaled by the age factor itself, which could then be used as an indicator to determine whether including temporal information is necessary to estimate the PageRank score of a node.

4.6. Rescaling PageRank's scores

A prediction of Eq. (24) is that, on average, we expect the age-rescaled PageRank score of a node to be proportional to its age-rescaled indegree. Numeric results on the APS citation network confirm that the rescaling procedure is effective at suppressing the age bias of both indegree and PageRank: For both variables, the score fold increase is close to one for all age bins (Fig 3A). When rescaling both PageRank score and indegree, the curves that describe the dependence of PageRank score on indegree for different age classes (Fig. 3B) collapse on top of each other, and for sufficiently large rescaled indegree scores, their behavior well agrees with direct proportionality (Fig 3B). The

rescaled indegree explains most of the variance of the rescaled PageRank score ($R^2 = 0.80$). At the same time, this correlation is lower than that in degree-uncorrelated networks, as reported below.

4.7. Decorrelating the APS network

Although the analytic results have been derived under the assumption of absence of degree-degree correlation, like in prior works [24, 25], the APS citation network exhibits degree-degree correlations (see [28] for a detailed analysis). Despite this mismatch, the analytic results accurately capture essential features of PageRank's behavior, including the moderating role of the age factor on its dependence on indegree, its stronger bias by age than indegree, and the effects of rescaling. Yet, the question remains whether the agreement between analytic and empirical results would improve if we reduced or eliminated the degree-degree correlations of the network.

To address this question, we extend all the previous results to degree-uncorrelated networks obtained from the original APS citation network. To obtain degree-uncorrelated networks from the original APS network while preserving the temporal structure of the data, we adopt a similar time-aware randomization procedure as that developed in [28]. In particular, within each calendar year y, we associate $\Delta k^*_{in}(i, y)$ and $\Delta k^*_{out}(i, y)$ stubs to each node i, where $\Delta k^*_{in}(i, y)$ and $\Delta k^*_{out}(i, y)$ denote the empirical variation of node i's indegree and outdegree, respectively, within year y. We then randomly match the in- and out-stubs, and eliminate self-loops and multiple edges [28]. Therefore, we end up with randomized networks where the nodes' indegree and outdegree values as well as their temporal evolution are approximately preserved. For simplicity, we discuss below the results for one single randomized network; the results for other realizations of the randomization procedure are qualitatively unchanged.

We find that the conclusions obtained in the APS empirical network still hold for the randomized network, although with a better match between analytic and empirical results. When considering the indegree predictor (Eq. (6)), the coefficient of determination between the nodes' predicted and empirical PageRank grows from $R^2 = 0.67$ (empirical network) to $R^2 = 0.75$ (randomized network). The coefficient of determination increases for the age-corrected predictor (Eq. (19)) as well: From $R^2 = 0.76$ (empirical network) to $R^2 = 0.86$ (randomized network). Therefore, in both the randomized and the empirical network, the age-corrected predictor enables a more accurate estimation of node PageRank than the indegree predictor. Similarly, the coefficient of determination between the nodes' predicted and empirical rescaled PageRank grows from $R^2 = 0.80$ (empirical) to $R^2 = 0.92$ (randomized network). Overall, these findings indicate that the discrepancies between the PageRank score and its expected value based on node age and indegree decrease after artificially suppressing the degree-degree correlations of empirical networks. In other words, in the randomized networks, the PageRank score is more accurately predictable from indegree and temporal patterns. These findings suggest that the added value of (rescaled) PageRank compared to (rescaled) indegree is greater in more correlated networks, which might explain its good performance at detecting significant papers and patents [21].

5. Results on synthetic networks

The results on empirical data call for a more systematic investigation of the accuracy of the age-corrected and indegree predictors and their dependence on the degree-degree correlations of the network. To this end, in this Section, we analyze synthetic data generated with a growing network model with a tunable degree of degree-degree correlation, which generalizes to directed networks a model previously proposed for undirected networks [31].

5.1. Network generation model

We describe here the network generation model and its control parameters. The proposed model generalizes the model by Guo et al. [31], proposed for undirected networks, to directed networks. The network growth begins by gradually introducing m initial nodes. Among these m initial nodes, the n-th one is connected with a directed link to all the previous $n - 1$ nodes. The remaining $N - m$ nodes are introduced one by one in time steps $m + 1, \ldots, N$; node n is thus added at time n. Each node forms m outgoing links. The target of the first link is chosen using preferential attachment (PA). After a link has been formed using the preferential attachment rule, with probability p the next link is formed by a triangle closing (TC) rule; otherwise, the next link is again formed using the PA rule ($p = 0$ thus recovers the standard preferential attachment model). The formation of links from node n stops when m outgoing links have been formed. The two mechanisms, PA and TC, are realized as follows:

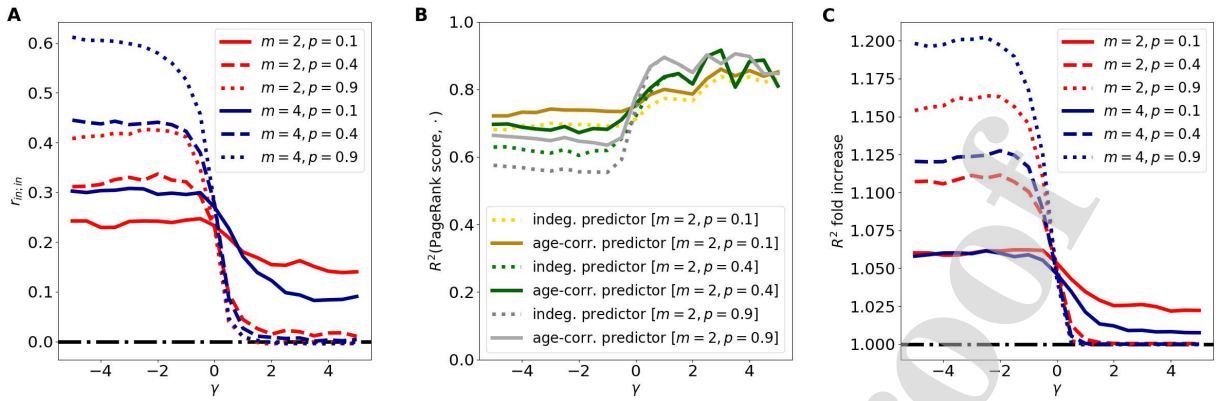


Figure 4: **Results in synthetic networks.** (A) Indegree-indegree assortativity coefficient as a function of γ. The correlation exhibits a sharp decline around $\gamma \simeq 0$, which marks a transition from a high-correlation regime ($\gamma < 1$) to a low-correlation regime ($\gamma > 1$). (B) As the indegree-indegree correlation decreases, the R^2 between the predicted PageRank score and the empirical one (represented on the y-axis) tends to increase for both the indegree and the age-corrected predictor. (C) Ratio between the R^2 of the age-corrected and the R^2 of the indegree predictor (R^2 "fold increase") as a function of γ. The R^2 fold increase is considerably larger in the high correlation regime.

- PA: If there is not yet a link from n to i, then the probability of i being chosen as the target of a given link from n is proportional to $k_{in}(i, n) + m$.
- TC: Denoting the target of the previous PA-driven link as i, triangle closing forms a link from n to a neighbor of i. If there is not yet a link from n to the neighbor j, then the probability of j being chosen is proportional to $[k_{in}(j, n) + m]^\gamma$.

The final network comprises N nodes. In our simulations, we use $N = 10,000$ and $m = 2$ or $m = 4$. All results presented below are averaged over 10 independent network realizations for each parameter setting.

By varying the model's control parameters, p and γ, we can tune the degree of assortativity of the resulting networks. We focus on the *indegree-indegree assortativity coefficient* of the generated networks, $r_{in;in}$, defined as the Pearson correlation coefficient between the indegree values found at the two extremes of the same directed edge [27]. Fig. 4A shows the behavior of $r_{in;in}$ as a function of γ for various values of p. We find that for negative γ values, the model generates networks with positive indegree-indegree correlations. Starting from negative γ values, as γ increases, we observe a transition to a regime of γ values where the generated networks exhibit small correlation values. For simplicity, we refer to the ranges $\gamma < 0$ and $\gamma > 0$ as the low- and high-correlation regime, respectively. When varying p, the dependence of the indegree-indegree correlation on γ remains qualitatively the same; at the same time, the larger p is, the more prominent the triangle closure mechanism is over the generation process, and the more assortative the resulting networks tend to be when $\gamma < 0$. Overall, the generated networks constitute an ideal playground to quantify the dependence of our analytic results for PageRank on the underlying network's correlations.

5.2. The role of assortativity

We find that the indegree-indegree assortativity significantly impacts the relative accuracy of the indegree and time-corrected predictor. In the low-correlation regime, we expect a high level of accuracy of both predictors as they both rely on the assumption of negligible degree-degree correlations. In this regime, both predictors achieve indeed high accuracy in estimating the PageRank score (Fig. 4B), and the age-corrected predictor exhibits marginally better accuracy than the indegree predictor (Fig. 4C). In the high-correlation regime, the assumption of negligible degree-degree correlations is violated. Predictably, in this regime, the accuracy of both the indegree and the age-corrected predictor are substantially lower than in the low-correlation regime, although the respective R^2 coefficients remain significantly larger than zero (Fig. 4B). Yet, it is in this regime that the age-corrected predictor achieves the largest accuracy gains compared to the indegree predictor (Fig. 4C).

Taken together, these results indicate that the age correction is particularly beneficial when the assumption of negligible degree-degree correlations is not supported. As indegree-indegree correlations increase, the accuracy of both predictors decreases, and the relatative accuracy of the age-corrected predictor compared to the indegree predictor improves. In other words, the benefits from taking into account temporal information to estimate the PageRank score are higher for correlated networks.

6. Conclusions

In this article, we aimed to understand analytically the emergence of a well-documented bias exhibited by the widely-used PageRank algorithm: Its bias by node age. Our findings demonstrate that in time-evolving directed networks, both indegree and temporal patterns play a key role in determining the PageRank score of a node. Using a mean-field approximation, we found that the PageRank score of a node can be expressed as the product of the node's indegree and an age factor that depends on the temporal structure of the network data. The age factor may "amplify" the indegree's age bias. When a significant amplification is found, such as in the analyzed APS citation network, the derived "age-corrected" predictor can substantially improve estimates of PageRank score compared to the previously-studied indegree predictor [24]. The same is not true in the absence of age bias amplification, as found in the analyzed Higgs Twitter dataset. Evidence from synthetic data shows that in more assortative networks, the accuracy of both predictors is lower, and the accuracy gain from the age-corrected predictor over the indegree predictor is substantially higher.

Several extensions of this work are possible. First, although we focused on PageRank's age bias, our analytic results could be extended to other biases found for the algorithm, e.g., its bias by scientific field in papers' citation networks [18]. Taking into account more potentially confounding factors could increase the accuracy of our estimates of a node's PageRank. Second, although our predictor explains most of the variance of PageRank scores, the correlation is not perfect. We would expect that the amount of variance that is not explained is partially due to what has made PageRank a successful algorithm: Its ability to identify nodes that are connected with other important nodes [7, 13]. Our analytic results could be used to develop a ranking algorithm that factors out the contributions of indegree and age from a node's score, which could help to detect low-degree nodes with high potential. Third, the proposed mean-field approximation for degree-age classes could be applied to understand the behavior of other network-based ranking algorithms in evolving networks. Algorithms that could be analyzed with our framework include those based on scores that can be written explicitly as functions of the network's adjacency matrix, e.g., the H-index [32], the collective influence algorithm [33], and PageRank variants with a non-uniform teleportation vector [12]. Fourth, while we have measured the temporal transition matrix $P(t' \to t)$ from empirical data (see Fig. 1), one could attempt to derive it analytically from mechanistic models of network formation. This could further deepen our understanding of how algorithmic behavior emerges from the microscopic behavior of the network's nodes. Finally, because of the increasing interest in ranking algorithms for multilayer [34] and higher-order networks [35], it could be beneficial to uncover potential biases of ranking algorithms based on more complex network representations.

Acknowledgements

This work was supported by the Fundamental Research Funds for the Central Universities, Southwest Minzu University (Grant No. 2020NZD02), Sichuan Science and Technology Program (Grant No. 2021YFH0120), Chengdu Science and Technology Program (Grant No. 2021-GH03-00001-HZ) and the National Natural Science Foundation of China (Grant No. 12050410248). MSM acknowledges financial support from the URPP Social Networks at the University of Zurich.

[1] E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, et al., Bias in data-driven artificial intelligence systems – An introductory survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (3) (2020) e1356.

[2] S. Zhang, M. Medo, L. Lü, M. S. Mariani, The long-term impact of ranking algorithms in growing networks, Information Sciences 488 (2019) 257–271.

[3] A. Sîrbu, D. Pedreschi, F. Giannotti, J. Kertész, Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model, PLOS ONE 14 (3) (2019) e0213246.

[4] P. Voosen, The AI detectives, Science 357 (2017) 22–27.

[5] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, et al., Machine behaviour, Nature 568 (7753) (2019) 477–486.
[6] N. E. Friedkin, Theoretical foundations for centrality measures, American Journal of Sociology 96 (6) (1991) 1478–1504.
[7] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems 30 (1) (1998) 107–117.
[8] A. N. Langville, C. D. Meyer, Google's PageRank and beyond: The science of search engine rankings, Princeton University Press, 2011.
[9] P. Chen, H. Xie, S. Maslov, S. Redner, Finding scientific gems with Google's PageRank algorithm, Journal of Informetrics 1 (1) (2007) 8–15.
[10] M. Dunaiski, J. Geldenhuys, W. Visser, Author ranking evaluation at scale, Journal of Informetrics 12 (3) (2018) 679–702.
[11] B. Jiang, S. Zhao, J. Yin, Self-organized natural roads for predicting traffic flow: A sensitivity study, Journal of Statistical Mechanics: Theory and Experiment 2008 (07) (2008) P07008.
[12] D. F. Gleich, PageRank beyond the Web, SIAM Review 57 (3) (2015) 321–363.
[13] M. Franceschet, PageRank: Standing on the shoulders of giants, Communications of the ACM 54 (6) (2011) 92–101.
[14] D. Walker, H. Xie, K.-K. Yan, S. Maslov, Ranking scientific publications using a model of network traffic, Journal of Statistical Mechanics: Theory and Experiment 2007 (06) (2007) P06010.
[15] M. S. Mariani, M. Medo, Y.-C. Zhang, Ranking nodes in growing networks: When PageRank fails, Scientific Reports 5 (2015) 16181.
[16] H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, M.-Y. Zhou, Ranking in evolving complex networks, Physics Reports 689 (2017) 1–54.
[17] F. Karimi, M. Génois, C. Wagner, P. Singer, M. Strohmaier, Homophily influences ranking of minorities in social networks, Scientific Reports 8 (1) (2018) 1–12.
[18] G. Vaccario, M. Medo, N. Wider, M. S. Mariani, Quantifying and suppressing ranking bias in a large citation network, Journal of Informetrics 11 (3) (2017) 766–782.
[19] M. S. Mariani, M. Medo, Y.-C. Zhang, Identification of milestone papers through time-balanced network centrality, Journal of Informetrics 10 (4) (2016) 1207–1223.
[20] M. S. Mariani, M. Medo, F. Lafond, Early identification of important patents: Design and validation of citation network metrics, Technological Forecasting and Social Change 146 (2019) 644–654.
[21] S. Xu, M. S. Mariani, L. Lü, M. Medo, Unbiased evaluation of ranking metrics reveals consistent performance in science and technology citation data, Journal of Informetrics 14 (1) (2020) 101005.
[22] M. S. Mariani, L. Lü, Network-based ranking in social systems: Three challenges, Journal of Physics: Complexity 1 (1) (2020) 011001.
[23] P. Holme, J. Saramäki, Temporal networks, Physics Reports 519 (3) (2012) 97–125.
[24] S. Fortunato, M. Boguñá, A. Flammini, F. Menczer, Approximating PageRank from in-degree, in: International Workshop on Algorithms and Models for the Web-Graph, Springer, 2006, pp. 59–71.
[25] S. Fortunato, M. Boguná, A. Flammini, F. Menczer, On local estimations of PageRank: A mean field approach, Internet Mathematics 4 (2-3) (2007) 245–266.
[26] M. De Domenico, A. Lima, P. Mougel, M. Musolesi, The anatomy of a scientific rumor, Scientific Reports 3 (1) (2013) 1–9.
[27] A.-L. Barabási, et al., Network science, Cambridge University Press, 2016.
[28] Z.-M. Ren, M. S. Mariani, Y.-C. Zhang, M. Medo, Randomizing growing networks with a time-respecting null model, Physical Review E 97 (5) (2018) 052311.
[29] A. Zeng, Z. Shen, J. Zhou, J. Wu, Y. Fan, Y. Wang, H. E. Stanley, The science of science: From the perspective of complex systems, Physics Reports 714 (2017) 1–73.
[30] F. Riquelme, P. González-Cantergiani, Measuring user influence on Twitter: A survey, Information Processing & Management 52 (5) (2016) 949–975.
[31] Q. Guo, T. Zhou, J.-G. Liu, W.-J. Bai, B.-H. Wang, M. Zhao, Growing scale-free small-world networks with tunable assortative coefficient, Physica A: Statistical Mechanics and its Applications 371 (2) (2006) 814–822.
[32] L. Lü, T. Zhou, Q.-M. Zhang, H. E. Stanley, The H-index of a network node and its relation to degree and coreness, Nature Communications 7 (1) (2016) 1–7.
[33] F. Morone, H. A. Makse, Influence maximization in complex networks through optimal percolation, Nature 524 (7563) (2015) 65–68.
[34] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, A. Arenas, Ranking in interconnected multilayer networks reveals versatile nodes, Nature Communications 6 (1) (2015) 1–6.
[35] F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, G. Petri, Networks beyond pairwise interactions: Structure and dynamics, Physics Reports.

1) In evolving networks, PageRank can exhibit stronger age biases than indegree
2) We use a mean-field approximation to understand the origin of PageRank's age bias
3) We estimate the expected PageRank score of nodes of a given indegree and age.
4) The obtained estimate is more accurate than estimates based on indegree alone.
5) We validate our results on empirical data and a growing network model with varying assortativity.

Mengtian Cui: Conceptualization; Analytic analysis; Roles/Writing - original draft.
Manuel S. Mariani: Conceptualization; Analytic analysis; Numeric analysis; Roles/Writing - original draft; Writing - review & editing.
Matus Medo: Conceptualization; Data curation; Supervision; Writing - review & editing.

Declaration of interests

■ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: