Journal of Informetrics 10 (2016) 1207–1223

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

Identification of milestone papers through time-balanced network centrality

Manuel Sebastian Mariani[a,*], Matúš Medo[a], Yi-Cheng Zhang[a,b]

[a] Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland
[b] Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, Chongqing 400714, P.R. China

ARTICLE INFO

Article history:
Received 18 June 2016
Received in revised form 14 October 2016
Accepted 14 October 2016
Available online 4 November 2016

Keywords:
Citation analysis
Metrics evaluation
PageRank
Time bias
Milestone papers

ABSTRACT

Citations between scientific papers and related bibliometric indices, such as the *h*-index for authors and the impact factor for journals, are being increasingly used – often in controversial ways – as quantitative tools for research evaluation. Yet, a fundamental research question remains still open: to which extent do quantitative metrics capture the significance of scientific works? We analyze the network of citations among the 449,935 papers published by the American Physical Society (APS) journals between 1893 and 2009, and focus on the comparison of metrics built on the citation count with network-based metrics. We contrast five article-level metrics with respect to the rankings that they assign to a set of fundamental papers, called Milestone Letters, carefully selected by the APS editors for “making long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research”. A new metric, which combines PageRank centrality with the explicit requirement that paper score is not biased by paper age, is the best-performing metric overall in identifying the Milestone Letters. The lack of time bias in the new metric makes it also possible to use it to compare papers of different age on the same scale. We find that network-based metrics identify the Milestone Letters better than metrics based on the citation count, which suggests that the structure of the citation network contains information that can be used to improve the ranking of scientific publications. The methods and results presented here are relevant for all evolving systems where network centrality metrics are applied, for example the World Wide Web and online social networks. An interactive Web platform where it is possible to view the ranking of the APS papers by rescaled PageRank is available at the address http://www.sciencenow.info.

1. Introduction

The notion of quantitative evaluation of scientific impact builds on the basic idea that the scientific merits of papers (Narin, 1976; Radicchi, Fortunato, & Castellano, 2008), scholars (Egghe, 2006; Hirsch, 2005), journals (Bollen, Rodriquez, & Van de Sompel, 2006; Liebowitz & Palmer, 1984; Pinski & Narin, 1976), universities (Kinney, 2007; Molinari & Molinari, 2008) and countries (Cimini, Gabrielli, & Labini, 2014; King, 2004) can be gauged by metrics based on the received citations. The respective field, referred to as bibliometrics or scientometrics, is undergoing a rapid growth (Van Noorden, 2010) fueled by the increasing availability of massive citation datasets collected by both academic journals and online platforms, such as

[*] Corresponding author.
E-mail address: manuel.mariani@unifr.ch (M.S. Mariani).

http://dx.doi.org/10.1016/j.joi.2016.10.005
1751-1577/© 2016 Elsevier Ltd. All rights reserved.

Google Scholar and Web of Science. The possible benefits, drawbacks and long-term effects of the use of bibliometric indices are being highly debated by scholars from diverse fields (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015; Lawrence, 2008; Van Raan, 2005; Weingart, 2005; Werner, 2015).

Although some effort has been devoted to contrast different metrics with respect to their ability to single out seminal papers (Dunaiski & Visser, 2012; Dunaiski, Visser, & Geldenhuys, 2016; Yao, Wei, Zeng, Fan, & Di, 2014; Zhou, Zeng, Fan, & Di, 2015), differences among the adopted benchmarking procedures and diverse conclusions of the mentioned references leave a fundamental question still open: which metric of scientific impact best agrees with expert-based perception of significance? In agreement with Wasserman, Zeng, and Amaral (2015), the significance of a scientific work is intended here as its enduring importance within the scientific community.

To address this question, we focus on a list of 87 physics papers of outstanding significance – called Milestone Letters – recently made available by the American Physical Society (APS) [http://journals.aps.org/prl/50years/milestones, accessed 25-11-2015]. According to the APS editors' description, the Milestone Letters "have made long-lived contributions to physics, either by announcing significant discoveries, or by initiating new areas of research". These articles have been carefully selected by the editors of the APS, and the choices are motivated in detail in the webpage; the fact that a large fraction of them led to Nobel Prize for some of their authors is an indication of the exceptional level of the selected works.

In this work, we analyze the network of citations between the $N = 449{,}935$ papers published in APS journals from 1893 until 2009 to compare five article-level metrics with respect to the ranking position they assign to the Milestone Letters. A reliable expert-based evaluation of the significance (intended as enduring importance, as in Wasserman et al., 2015) of a paper necessarily requires a time lag between the paper's publication date and the expert's judgment. For example, there is a time interval of 14 years between the most recent Milestone Letter (from 2001) and the year at which the list of Milestone Letters was released (2015). However, we show that a well-designed quantitative metric offers us the opportunity to detect potentially significant papers relatively short after their publication – an aspect often neglected in the evaluation of bibliometric indicators. To show this, we study how the ability of the different metrics to identify the Milestone Letters changes with paper age.

A plethora of quantitative metrics exist and could be studied in principle. Our focus here is narrowed to metrics that rely on a diffusion process on the underlying network of citations between papers and their comparison with simple citation count. The five metrics considered in this work are thus: the citation count, PageRank (introduced by Brin & Page, 1998), CiteRank (introduced by Walker, Xie, Yan, & Maslov, 2007), rescaled citation count (introduced by Newman, 2009), and novel rescaled PageRank. PageRank is a classical network centrality metric which combines a random walk along network links with a random teleportation process. The metric has been applied to a broad range of real-world problems (Ermann, Frahm, & Shepelyansky, 2015; Franceschet, 2011; Gleich, 2015 for a review), including ranking academic papers (Chen, Xie, Maslov, & Redner, 2007; Yao et al., 2014), journals (Bollen et al., 2006; González-Pereira, Guerrero-Bote, & Moya-Anegón, 2010) and authors (Nykl, Ježek, Fiala, & Dostal, 2014; Radicchi, Fortunato, Markines, & Vespignani, 2009; Yan & Ding, 2009) (see Waltman & Yan, 2014 for a review of the applications of PageRank-related methods to bibliometric analysis).

To overcome the well-known PageRank's bias toward old nodes in citation data (detailedly studied by Chen et al., 2007; Mariani, Medo, & Zhang, 2015), the CiteRank algorithm introduces exponential penalization of old nodes, resulting in a node score that well captures the future citation count increase of the papers and, for this reason, can be considered as a reasonable proxy for network traffic, as shown by Walker et al. (2007). However, we show below that CiteRank score does not allow one to fairly compare papers of different age. Rescaled citation count and rescaled PageRank are derived from citation count and PageRank score, respectively, by explicitly requiring that paper score is not biased by age – the adopted rescaling procedure is conceptually close to the methods recently developed by Radicchi et al. (2008), Newman (2009), Radicchi and Castellano (2011), Newman (2014), Radicchi and Castellano (2012b), Radicchi and Castellano (2012a), Crespo, Ortuño-Ortín, and Ruiz-Castillo (2012) and Kaur, Ferrara, Menczer, Flammini, and Radicchi (2015) to suppress biases by age and field in the evaluation of academic agents. We find that the rankings produced by the rescaled scores are indeed consistent with the hypothesis that the rankings are not biased by age.

We find that PageRank can compete and even outperform rescaled PageRank in identifying *old* milestone papers, but completely fails to identify recent milestone papers due to its temporal bias. CiteRank can compete and even outperform rescaled PageRank in identifying *recent* milestone papers, but markedly underperforms in identifying old milestone papers due to its built-in exponential penalization for older papers. Indicators based on simple citation count are outperformed by rescaled PageRank for papers of every age. This leads us to the conclusion that rescaled PageRank is the best-performing metric overall. With respect to previous works by Chen et al. (2007), Dunaiski and Visser (2012), Fiala (2012) and Dunaiski et al. (2016) that claimed the superiority of network-based metrics in identifying important papers, our results clarify the essential role of paper age in determining the metrics' performance: rescaled PageRank excels and PageRank performs poorly in identifying MLs short after their publication, and the performance of the two methods becomes comparable only 15 years after the MLs are published. Qualitatively similar results are found for an alternative list of APS outstanding papers which only includes works that have led to Nobel prize for some of the authors (the list is provided in the Table S2).

Our results indicate that network centrality and time-balance are two essential ingredients – though neglected by popular bibliometric indicators such as the *h*-index for scholars (Hirsch, 2005) and impact factor for journals (Garfield, 1972) – for an effective detection of significant papers. This sets a new benchmark for article-level metrics and quantitatively support the paradigm that considering the whole network instead of simple citation count can bring substantial benefits to the ranking of academic agents. In a broader context, our results show that a direct rescaling of PageRank scores is an effective method to solve the PageRank's well-known bias against recent network nodes. We emphasize that while scientific papers are the focus of this work, the addressed research question is general and can emerge when estimating the importance of any creative work – such as movies (Spitz & Horvát, 2014; Wasserman et al., 2015) – for which quantitative impact metrics and expert-based significance assessments are simultaneously available. The potential broader applications and possible limitations of our results are discussed in the Discussion section.

2. Metrics

We consider five article-level metrics: citation count c, PageRank score p, CiteRank score T, rescaled PageRank score $R(p)$, and rescaled citation count $R(c)$.

2.1. Citation count

We denote by $\mathbf{A}$ the network's adjacency matrix (A_{ij} is one if node j points to node i and zero otherwise). Citation count (referred to as indegree in network science literature, see Newman, 2010) is one of the simplest metrics to infer node centrality in a network, being simply defined as $c_i = \sum_j A_{ij}$ for a node i. Citation count is the building block of the majority of metrics for assessing the impact of single papers, authors, journals (for a review of citation-based impact indicators see Waltman, 2016).

2.2. PageRank

The PageRank score vector was introduced by Brin and Page (1998), and can be defined as the stationary state of a process which combines a random walk along the network links and random teleportation. In a directed monopartite network composed of N nodes, the vector of PageRank scores $\{p_i\}$ can be found as the stationary solution of the following set of recursive linear equations

$$p_i^{(n+1)} = \alpha \sum_{j:k_j^{out}>0} A_{ij} \frac{p_j^{(n)}}{k_j^{out}} + \alpha \sum_{j:k_j^{out}=0} \frac{p_j^{(n)}}{N} + \frac{1-\alpha}{N}, \tag{1}$$

where $k_j^{out} := \sum_l A_{lj}$ is the outdegree of node j, α is the teleportation parameter, and n is the iteration number. Eq. (1) represents the master equation of a diffusion process on the network, which converges to a unique stationary state independently of the initial condition (see Berkhin, 2005 for the mathematical details). The PageRank score p_i of node i can be interpreted as the average fraction of time spent on node i by a random walker who with probability α follows the network's links and with probability $1-\alpha$ teleports to a random node. Throughout this paper, we set $\alpha = 0.5$ which is the usual choice for citation data (Chen et al., 2007).

2.3. CiteRank

To correct the PageRank's strong temporal bias in citation networks, the CiteRank algorithm (introduced by Walker et al., 2007) introduces ad hoc penalization for older nodes. The CiteRank score T is defined similarly as PageRank; differently from PageRank, in CiteRank equations the teleportation probability decays exponentially with paper age with a certain timescale τ. According to Walker et al. (2007) and Maslov and Redner (2008), this choice of the teleportation vector is intended to favor the recent nodes and thus lead to a score that better represents papers' relevance for the current lines of research. Using the same notation as Eq. (1), the vector of CiteRank scores $\{T_i\}$ can be found as the stationary solution of the following set of recursive linear equations

$$T_i^{(n+1)} = \alpha \sum_{j:k_j^{out}>0} A_{ji} \frac{T_j^{(n)}}{k_j^{out}} + \alpha \sum_{j:k_j^{out}=0} \frac{T_j^{(n)}}{N} + (1-\alpha) \frac{\exp(-(t-t_i)/\tau)}{\sum_{j=1}^{N} \exp(-(t-t_j)/\tau)}, \tag{2}$$

where we denote by t_i the publication date of paper i and t the time at which the scores are computed. Throughout this paper we set $\alpha = 0.5$ and $\tau = 2.6$ years, which are the parameters chosen by Walker et al. (2007). The performance of the algorithm for other values of the parameter τ is discussed in the caption of Fig. E.10, in Appendix E. We show below that exponential penalization of older nodes is not effective in removing PageRank's bias, and propose instead a rescaled PageRank score $R(p)$ whose average value and standard deviation do not depend on paper age.

2.4. Rescaled PageRank and rescaled citation count

To compute the rescaled PageRank score $R(p)$ for a given paper i, we evaluate the paper's PageRank score p_i as well as the mean $\mu_i(p)$ and standard deviation $\sigma_i(p)$ of PageRank score for papers published in a similar time as i. Time is not measured in days or years, but in number n of published papers; after labeling the papers in order of decreasing age, $\mu_i(p)$ and $\sigma_i(p)$ are computed over papers $j \in [i - \Delta_p/2, i + \Delta_p/2]$. The parameter Δ_p represents the number of papers in the averaging window of each paper.[1] The rescaled score $R_i(p)$ of paper i is then computed as

$$R_i(p) = \frac{p_i - \mu_i(p)}{\sigma_i(p)}. \tag{3}$$

Values of $R(p)$ larger or smaller than zero indicate whether the paper is out- or under-performing, respectively, with respect to papers of similar age. $R_i(p)$ represents the z-score (Kreyszig, 2010) of paper i within its averaging window. For the sake of completeness, we have also tested a simpler rescaled score in the form $R_i^{(ratio)}(p) = p_i/\mu_i(p)$; however, $R^{(ratio)}(p)$ fails to produce a time-balanced ranking due to the fact that $\sigma(p)/\mu(p)$ strongly depends on paper age (see Appendix C for details). In addition, we tested a rescaled score $R^{(year)}(p)$ based on Eq. (3) where $\mu_i(p)$ and $\sigma_i(p)$ are computed over the papers published in the same year as paper i. We found that while $R^{(year)}(p)$ is able to suppress large part of PageRank's temporal bias, its ranking is much less in agreement with the hypothesis of unbiased ranking than the ranking by $R(p)$ (see Appendix C for details). For this reason, we use an averaging window based on number of publications and not on real time. This choice is also supported by the findings by Newman (2009) and Parolo et al. (2015) which suggest that the role of time in citation networks is better captured by the number of published papers than by real time.

We define the rescaled citation count analogously as

$$R_i(c) = \frac{c_i - \mu_i(c)}{\sigma_i(c)}, \tag{4}$$

where $\mu_i(c)$ and $\sigma_i(c)$ represent the mean and the standard deviation of c computed over papers $j \in [i - \Delta_c/2, i + \Delta_c/2]$. Citation count rescaling was used by Newman (2009) and Newman (2014) to identify papers that accrue more citations than expected for papers of similar age under the hypothesis of pure preferential attachment.

The choice of the size of the temporal window deserves some attention: if the size of the temporal window is too large, one would fall again in a time-biased ranking that is one of the issues that motivate the present paper. On the other hand, if we choose a too small averaging window, the papers would be only compared with few other papers and the resulting scores would be too volatile. Throughout this paper, we set $\Delta_c = \Delta_p = 1000$; we refer to Appendix D for further details on the dependence of ranking properties on the averaging window size. We stress that the rankings by $R(c)$ and $R(p)$ are only weakly dependent on Δ_c and Δ_p (see Fig. D.9), and the correlation between the rankings by $R(p)$ obtained with different values of PageRank's teleportation parameter α is close to one (Spearman's rank correlation coefficient between the rankings obtained with $\alpha = 0.5$ and $\alpha = 0.85$ is equal to 0.98). These results indicate that the proposed rescaling metrics are robust with respect to variations of their parameters.

3. Results

We analyzed the network composed of $L = 4{,}672{,}812$ citations among $N = 449{,}935$ papers published in APS journals (1893–2009). The dataset was directly provided by the APS following our request at the webpage http://journals.aps.org/datasets, and was also studied by, among others, Medo, Cimini, and Gualdi (2011).

3.1. Time balance of the rankings

Before comparing the performances of the five metrics in recognizing the Milestone Letters (MLs), we want to determine whether the metrics are biased by age and, if yes, then to which extent. In agreement with Radicchi et al. (2008) and Radicchi and Castellano (2011), we assume that a fair ranking of scientific papers should be time-balanced in the sense that old and recent papers should be equally likely to appear at the top of the ranking by the metric. Caveats and possible weak points of this assumption are examined in the Discussion section.

To assess the degree of time balance of the five metrics, we perform a statistical test similar to those proposed by Radicchi and Castellano (2011) and Radicchi and Castellano (2012b). We divide the papers into $S = 40$ different groups according to their age and, for each metric, we compute the number $n_\alpha(z)$ of top-zN papers by the metric for each age group α, and quantitatively compare the resulting histogram $\{n_\alpha(z)\}$ with the expected histogram $\{n_\alpha^{(0)}(z)\}$ under the hypothesis that the ranking is temporally unbiased. We set $z = 0.01$; results for other small values of z are qualitatively similar.

Fig. 1A shows that the observed values of $n(0.01)$ for PageRank are far from their expected values under the hypothesis of unbiased ranking. For instance, $n_1(0.01)/n_1^{(0)}(0.01) = 4.62$ for the age group that contains the oldest $N/40$ papers, as opposed to $n_{40}(0.01) = 0$ for the age group composed of the most recent $N/40$ papers. To quantify the degree of time balance of a metric, we compare the standard deviation σ of the observed histogram $\{n_\alpha(0.01)\}$ with the expected standard deviation σ_0 under the hypothesis of unbiased ranking. For a perfectly unbiased ranking, the number $n_\alpha^{(0)}$ of nodes from age group α in the top-z

[1] In order to have the same number of papers in each averaging window, a different definition of averaging window is needed for the oldest and the most recent $\Delta_p/2$ papers, for which we compute μ_i and σ_i over the papers $j \in [1, \Delta_p]$ and $j \in (N - \Delta_p, N]$, respectively.

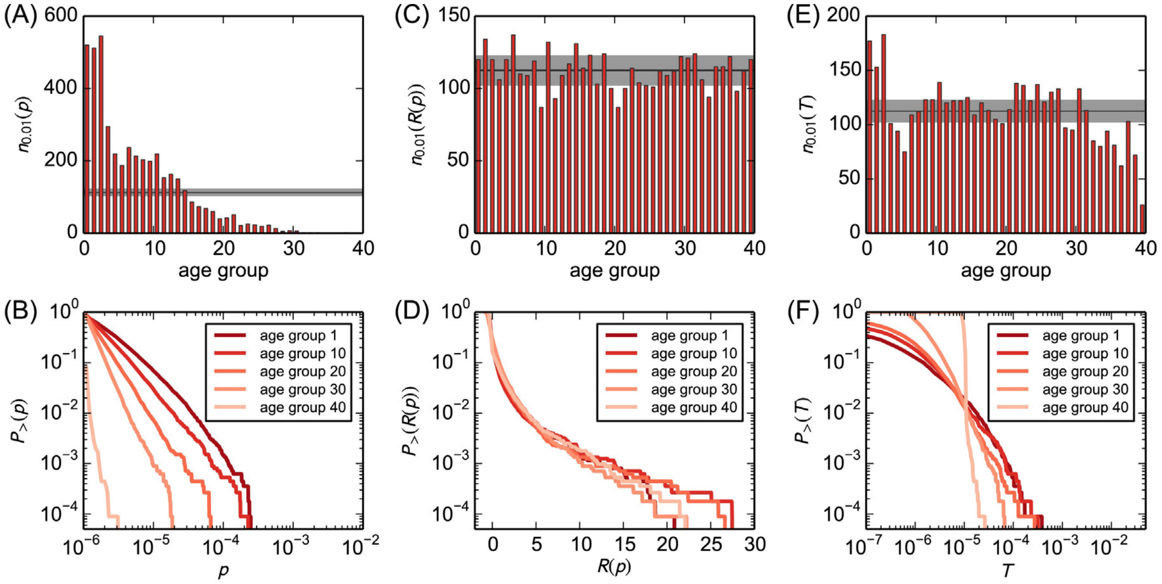

Fig. 1. Time balance of the network-based metrics. Panels (A, C, E) show the histogram of the number of papers from each paper age group in the top-1% of the ranking by PageRank score p, rescaled PageRank score $R(p)$ and CiteRank score T, respectively (age group 1 and age group 40 contain the oldest and most recent $N/40$ papers, respectively). The horizontal black line represents the unbiased value $n^{(0)}(0.01) = 0.01\,N/40$; the gray-shaded area represents the interval $[n^{(0)} - \sigma_0, n^{(0)} + \sigma_0]$ with σ_0 given by Eq. (5). Panels (B, D, F) show the cumulative distributions of PageRank score p, rescaled PageRank score $R(p)$ and CiteRank score T, respectively, for different age groups.

by the ranking obeys the multivariate hypergeometric distribution (Radicchi & Castellano, 2012b). Therefore, we expect on average $n^{(0)}(z) = zN/S$ top-zN papers for each set, with the standard deviation

$$\sigma_0(z) = \sqrt{\frac{zN}{S}\left(1-\frac{1}{S}\right)(1-z)\frac{N}{N-1}}, \tag{5}$$

The observed standard deviation $\sigma(z)$ is computed as

$$\sigma(z) = \sqrt{\frac{1}{S}\sum_{\alpha=1}^{S}\left(n_\alpha - n_\alpha^{(0)}\right)^2}. \tag{6}$$

The ratio σ/σ_0 between observed and expected standard deviation quantifies the degree of time balance of the ranking – we expect this ratio to be close to or lower than (due to fluctuations) one for an unbiased ranking, and significantly larger than one for a ranking biased by age. To quantify to which extent the observed values of $\sigma/\sigma_0 - 1$ are consistent with the hypothesis of unbiased ranking, we run a simulation where $0.01\,N$ papers are randomly assigned to one among 40 groups, and compute the standard deviation σ_{dev} of the observed deviation $\sigma_{rand}/\sigma_0 - 1$ according to Eq. (6). With 10^5 realizations, we obtain $\sigma_{dev} = 0.11$. We always express the observed values of $\sigma/\sigma_0 - 1$ as multiples of σ_{dev} in the following.

We obtain $\sigma/\sigma_0 - 1 = 12.91 = 117.36\,\sigma_{dev}$ for PageRank, which indicates that the ranking is heavily biased. The heavy bias of PageRank score is also revealed by a comparison of its distribution for nodes from different age groups, which shows a clear advantage for old nodes (Fig. 1B). Fig. 1C shows that the ranking by the $R(p)$ score is in good agreement with the hypothesis that the ranking is unbiased; we find $\sigma/\sigma_0 - 1 = 0.16 = 1.45\,\sigma_{dev}$. The time balance of rescaled PageRank score manifests itself in the collapse of the distributions of the $R(p)$ score for different age groups on a unique curve, which means that the $R(p)$ score allows us to compare papers of any age on the same scale (Fig. 1D). In a similar way, the rescaling procedure suppresses the temporal bias of citation count [$\sigma/\sigma_0 - 1 = 0.10 = 0.91\,\sigma_{dev}$ for $R(c)$ as compared to $\sigma/\sigma_0 - 1 = 6.01 = 54.64\,\sigma_{dev}$ for c, see Fig. 2]. We observe a qualitatively similar suppression of time bias for different choices of the number S of age groups (not shown here).

With respect to the histogram obtained with $R(p)$, the histogram $\{n_\alpha(0.01)\}$ obtained with the CiteRank algorithm (with the parameters chosen by Maslov and Redner (2008)) presents much larger deviation from the histogram expected under the hypothesis of time-balanced ranking (see Fig. 1E). As a result, the value of σ/σ_0 obtained for CiteRank ($\sigma/\sigma_0 - 1 = 1.75 = 15.91\,\sigma_{dev}$ with the parameters chosen by Walker et al. (2007)) is larger than the value obtained for $R(p)$. The distributions of CiteRank score T for different age groups do not collapse on a single curve (see Fig. 1F), which is directly due to the built-in exponential decay of the teleportation term. The failure of CiteRank in producing a time-balanced ranking is well exemplified by the behavior of the score distribution for the most recent age group, whose minimum score (i.e.,

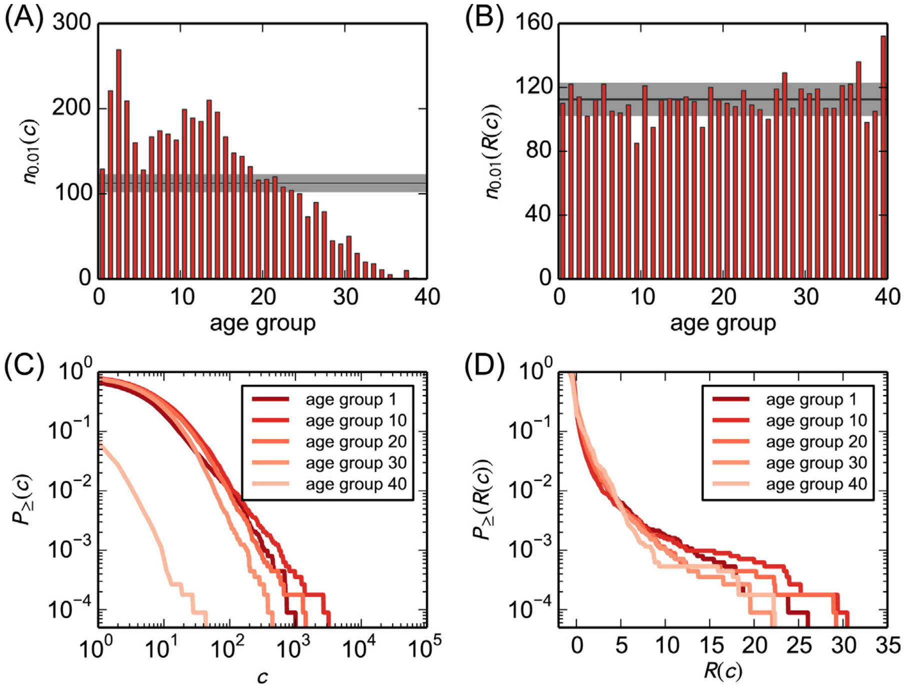

Fig. 2. Time balance of the citation-based metrics. Panels (A, B) show the histogram of the number of top-1% papers for each paper age group in the ranking by citation count c and rescaled citation count $R(c)$, respectively. Panels (C, D) show the cumulative distributions for different age groups of citation count c and rescaled citation count $R(c)$, respectively.

the smallest score value such that $P(>T)$ deviates from one) is much larger than for the other distributions, due to a larger teleportation term. These findings show that CiteRank score does not allow us to fairly compare papers of different age.

The values of $\sigma/\sigma_0 - 1$ for the five metrics are summarized in Table 1.

3.2. Identification of the Milestone Letters

In the previous section, we have shown that the rankings by the rescaled metrics $R(p)$ and $R(c)$ are consistent with the hypothesis that the ranking is not biased by paper age. While different works have recently emphasized the importance of removing the bias by age of citation-performance metrics for a fair ranking of scientific publications (Radicchi & Castellano, 2011; Radicchi et al., 2008) and researchers (Kaur, Radicchi, & Menczer, 2013), the possible positive effects of time-balanced rankings with respect to biased rankings remain largely unexplored.

Chen et al. (2007) analyzed the APS dataset and found that PageRank is able to recognize old papers that are universally important for physics. They also noted that PageRank is based on a diffusion process that drifts towards old papers (see Mariani et al., 2015 for a general analysis of this aspect) and, as a consequence, it inevitably favors old papers. Since the rescaling procedure that we propose solves this issue, it is thus plausible to conjecture that with respect to the PageRank algorithm, rescaled PageRank allows us to identify seminal papers earlier.

In this section, we use the APS dataset and the list of Milestone Letters (MLs) chosen by editors of Physical Review Letters (see Supplementary Table S1 for the list of MLs) to address the two following research questions:

Table 1
The five considered metrics and their bias by age. The difference $\sigma/\sigma_0 - 1$ quantifies how much the histogram of the number of top-1% papers by the metric deviates from the histogram expected under the hypothesis of ranking not biased by age (see the main text). The values of $\sigma/\sigma_0 - 1$ are expressed as multiples of their expected value $\sigma_{dev} = 0.11$ for a random ranking of the papers (computed as explained in the main text). Values of $\sigma/\sigma_0 - 1$ smaller than $2\,\sigma_{dev} = 0.22$ are reported in bold characters.

Metric	Properties	$\sigma/\sigma_0 - 1$
Citation count c	Local metric	$54.64\,\sigma_{dev}$
PageRank score p	Network-based metric	$117.36\,\sigma_{dev}$
CiteRank T	Network-based metric, time-aware	$15.91\,\sigma_{dev}$
Rescaled PageRank $R(p)$	Network-based metric, time-aware	**1.45** σ_{dev}
Rescaled citation count $R(c)$	Local metric, time-aware	**0.91** σ_{dev}

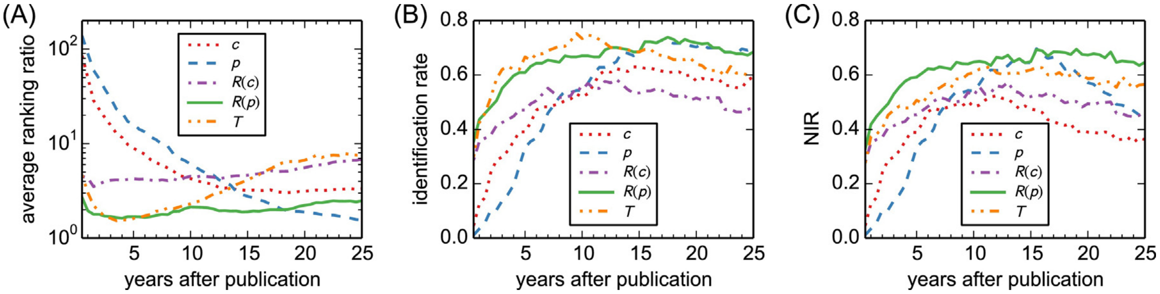

Fig. 3. Metrics' performance in ranking the milestone letters (listed in the Supplementary Table S1) as a function of paper age. (A) Dependence of the average ranking ratio $\bar{r}$ on paper age. (B) Dependence of the identification rate $f_{0.01}$ on paper age. (C) Dependence of the normalized identification rate $\tilde{f}_{0.01}$ on paper age.

1. *Is there a significant gap between the performance of rescaled PageRank and PageRank in identifying the MLs short after publication? If there is a substantial gap, does it close down after a certain number of years after publication?*
2. *Do network-based indicators outperform indicators based on simple citation count in recognizing the MLs?*

To compare the ranking positions of the MLs by the five different metrics, the ranking of Milestone Letter i is computed t years after its publication. We calculate the ratio of i's ranking position $r_i(s, t)$ by metric s and i's best ranking position $\min_s\{r_i(s, t)\}$ among all considered metrics. To characterize the overall performance of metric s in ranking the MLs, we average the ranking ratio over i and obtain $\bar{r}(s, t)$ (see F for computation details). The resulting quantity is referred to as the *average ranking ratio* of metric s for the Milestone Letters t years after their publication. A good metric is expected to have as low $\bar{r}(\cdot, t)$ as possible – the minimum value $\bar{r}(\cdot, t) = 1$ is only achieved by a metric that always outperforms the others in ranking the milestone papers of age t. Note that the average ranking ratio reduces to average ranking position if we do not normalize the ranking position $r_i(s, t)$ by $\min_s\{r_i(s, t)\}$. However, the average ranking position of the target papers by a certain metric is extremely sensitive to the ranking positions of the least-cited target papers, as opposed to the robustness of the average ranking ratio with respect to removal of the least-cited papers from the set of target papers (see Appendix A for details). This property motivates the use of ranking ratio to compare the ranking positions of the MLs by the different metrics.

The dependence of $\bar{r}(s, t)$ on paper age t measured in years after publication is shown in Fig. 3A. Due to the suppression of time bias, rescaled PageRank score $R(p)$ has a large advantage with respect to the original PageRank score p for papers of small age. Since the PageRank algorithm is biased towards old nodes, the performance gap between $R(p)$ and p gradually decreases with age and vanishes 18 years after publication. By contrast, the CiteRank algorithm exponentially penalizes older nodes and, as a consequence, the performance gap between $R(p)$ and T is minimal for recent papers, and CiteRank score T can even outperform $R(p)$ during the first six years after publication. When paper age becomes sufficiently larger than CiteRank's temporal timescale ($\tau = 2.6$ years here, as chosen by Walker et al. (2007) and Maslov and Redner (2008)), older papers are strongly penalized by the CiteRank's teleportation term and, as a result, CiteRank is markedly outperformed by rescaled PageRank. The same behavior is observed also for other values of CiteRank time-decay parameter τ (see Appendix E). The local metrics c and $R(c)$ are outperformed by $R(p)$ in ranking the MLs of every age, which indicates that network centrality brings a substantial advantage in ranking highly significant papers with respect to simple and rescaled citation count.

While the average ranking ratio $\bar{r}$ takes into account all the MLs, it is also interesting to measure the age-dependence of the identification rates of the metrics, defined as the fraction $f_x(t)$ of MLs that were ranked among the top $x\,N$ papers by the metric when they were t years old[2] (see Fig. 3B). Rescaled PageRank $R(p)$ and CiteRank score T markedly outperform the other metrics in identifying the milestone papers in the first years after publication. The performance gap between $R(p)$ and the citation-based indicators c and $R(c)$ remains significant during the whole observation lapse. Analogously to what we observed for the average ranking ratio, the performance gap between $R(p)$ and p gradually decreases with paper age and vanishes 15 years after publication, which is similar to the crossing point at 18 years after publication observed for the average ranking ratio. CiteRank has a small advantage with respect to rescaled PageRank in the first years after publication, whereas for older papers CiteRank's identification rate drops to the value achieved by simple citation count c.

It is worth to observe that the temporal bias of a certain metric affects the behavior of both $\bar{r}(t)$ and $f_{0.01}(t)$ for that metric: as we observe in Appendix B, a metric biased towards old (like PageRank) or recent papers naturally performs better in identifying old or recent MLs, respectively. One natural way to understand this effect is to consider a normalized identification rate $\tilde{f}_{0.01}(t)$ (hereafter abbreviated as NIR), such that the contribution of each identified ML i of age t (i.e., a ML ranked in the top $0.01\,N$ of the ranking) to $\tilde{f}_{0.01}(t)$ is smaller than one if the metric favors papers that belong to the same age group as paper i (see Appendix F for the mathematical definition). In other words, when evaluating the performance of a

[2] The identification rate is related to *recall*, a standard measure in the literature of recommendation systems (Lü et al., 2012).

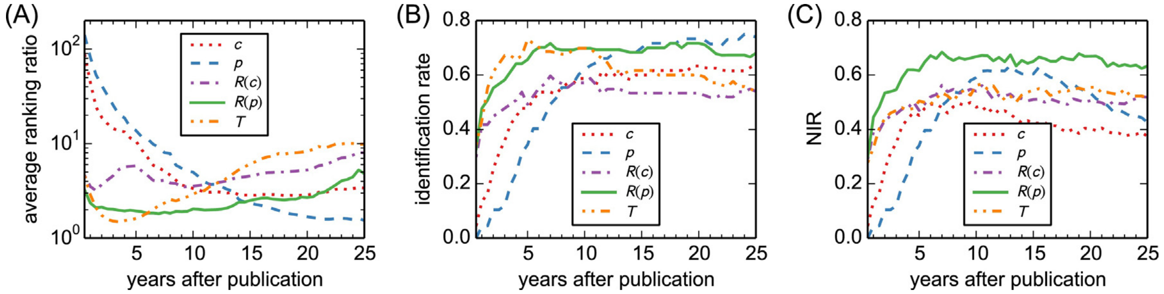

Fig. 4. Metrics' performance in ranking the APS papers that led to Nobel prize for some of the authors, listed in the Supplementary Table S2. The figure has been realized with the same procedure used for Fig. 3. (A) Dependence of the average ranking ratio $\bar{r}$ on paper age. (B) Dependence of the identification rate $f_{0.01}$ on paper age. (C) Dependence of the normalized identification rate $\tilde{f}_{0.01}$ on paper age. We observe a behavior in qualitative agreement with that observed in Fig. 3.

given metric, the normalized identification rate $\tilde{f}_{0.01}(t)$ takes into account both the temporal balance and the identification power of the metric. The behavior of $\tilde{f}_{0.01}(t)$ for the five metrics is shown in Fig. 3C. After an initial increasing trend for all the metrics, the normalized identification rate of both p and c decline due to their temporal bias; by contrast, the same quantity remains relatively stable for both $R(p)$ and $R(c)$. According to $\tilde{f}_{0.01}(t)$, rescaled PageRank outperforms CiteRank for papers of every age. This is due to the fact that the ranking by CiteRank is not unbiased and, as a consequence, CiteRank's performance is often penalized by the NIR for small age t due to the algorithm's bias towards recent nodes.

Our analysis assumes that a ML should be ranked as high as possible by a good metric for scientific significance. On the other hand, many outstanding contributions to physics are not included in the list of MLs. To show that our results also hold for an alternative choice of groundbreaking papers, we consider a list of 67 APS papers that led to Nobel Prize for some of the authors (see Supplementary Table S1 for the list of papers). The results for this list of benchmark papers are shown in Fig. 4 and are qualitatively similar to those shown in Fig. 3, which indicates that our findings are robust with respect to modifications of the benchmark papers' list.

While Fig. 3 concerns the metrics' performance averaged over the whole set of MLs, the Supplementary Movie shows the simultaneous dynamics of the ranking positions by p and $R(p)$ of all individual MLs for the first 15 years after publication.[3] The movie shows that rescaled score $R(p)$ has a clear advantage with respect to PageRank score p in the first years after publication for most of the MLs. As the MLs become sufficiently old, their position in the plane gradually tends to converge to the diagonal where the ranking position by p is equal to the ranking position by $R(p)$, which is in agreement with the crossing between PageRank's and rescaled PageRank's performance curves observed in Fig. 3A and B.

In principle, one might consider a comparison of the final ranking positions (i.e., the ranking positions computed on the whole dataset) of the target papers by a certain metric (Dunaiski et al., 2016; Dunaiski & Visser, 2012) instead of the age-dependent evaluation of the metrics introduced above. But this kind of comparison would miss our key point – the strong dependence of metrics' performance on paper age. In addition, the strong dependence of metrics' performance on paper age shown in this section makes the outcome of such evaluation strongly dependent on the age distribution of the target papers we aim to identify. This issue is discussed in Appendix B and potentially concerns any performance evaluation carried out on a fixed snapshot of the network. By contrast, the outcomes presented in this paragraph (how well do the different metrics perform as a function of paper age) are little sensitive to the exact age distribution of the target papers.

3.3. Top papers by PageRank and rescaled PageRank

To get an intuitive understanding of the properties of PageRank and its rescaled version, it is instructive to look at the top-15 papers according to p and $R(p)$ computed on the whole dataset, reported in Tables 2 and 3, respectively. Although only one ML appears in the top 15 by p (ranked 6th, see Table 2), among the non-MLs there are papers of exceptional significance, such as the letter that proposed the popular Einstein–Podolsky–Rosen experiment (ranked 7th); the paper that introduced a fundamental tool in many-body systems, Slater's determinant (ranked 5th); the paper that presented the famous exact solution of the two-dimensional Ising model (ranked 8th). This confirms that PageRank is highly effective in finding relatively old papers of outstanding significance – referred to as "scientific gems" by Chen et al. (2007) – which has led to the interpretation of PageRank score as a "lifetime achievement award" for a paper (Maslov & Redner, 2008). Nevertheless, the most recent paper in Table 2 is from 1981 – 28 years old with respect to the dataset's ending point in 2009.

In the top-15 by $R(p)$, we find both old papers (the oldest is from 1964, 45 years old in 2009) and recent papers (the most recent is from 2002, 7 years old in 2009). Four out of 15 top-papers are MLs, which is an additional confirmation of the quality of the ranking by $R(p)$. We emphasize that while both PageRank and rescaled PageRank feature prominent papers

[3] Accordingly, only the 73 MLs that are at least 15 years old at the end of the dataset are included in the movie.

Table 2
Top-15 papers in the APS data as ranked by PageRank score p (asterisks mark the Milestone Letters).

Rank (p)	Rank ($R(p)$)	$p(\times 10^{-5})$	$R(p)$	Title	Year	Journal
1	1	43.32	29.96	Self-consistent equations including exchange and correlation effects (W. Kohn, L. Sham)	1965	*Phys. Rev.*
2	36	40.77	24.57	Theory of superconductivity (J. Bardeen, L. Cooper, J. Schrieffer)	1957	*Phys. Rev.*
3	8	35.88	28.58	Inhomogeneous electron gas (P. Hohenberg)	1964	*Phys. Rev.*
4	115	24.74	18.64	Stochastic problems in physics and astronomy (S. Chandrasekhar)	1943	*Rev. Mod. Phys.*
5	137	23.57	17.78	The theory of complex spectra (J. Slater)	1929	*Phys. Rev.*
6	21	23.46	26.53	*A model of leptons (S. Weinberg)	1967	*Phys. Rev. Lett.*
7	130	22.80	18.05	Can quantum-mechanical description of physical reality be considered complete? (A. Einstein, B. Podolsky, N. Rosen)	1935	*Phys. Rev.*
8	140	22.67	17.73	Crystal statistics. I. A two-dimensional model with an order-disorder transition (L. Onsager)	1944	*Phys. Rev.*
9	15	22.64	27.44	Self-interaction correction to density-functional approximations for many-electron systems (J. Perdew)	1981	*Phys. Rev. B*
10	335	22.39	13.17	Absence of diffusion in certain random lattices (P. Anderson)	1958	*Phys. Rev.*
11	16	21.25	26.88	Scaling theory of localization: absence of quantum diffusion in two dimensions (E. Abrahams)	1979	*Phys. Rev. Lett.*
12	110	20.67	18.83	Effects of configuration interaction on intensities and phase shifts (U. Fano)	1961	*Phys. Rev.*
13	82	19.36	20.86	On the constitution of metallic sodium (E. Wigner, F. Seitz)	1933	*Phys. Rev.*
14	210	18.32	15.44	On the interaction of electrons in metals (E. Wigner)	1934	*Phys. Rev.*
15	315	18.25	13.53	Cohesion in monovalent metals (J. Slater)	1930	*Phys. Rev.*

Table 3
Top-15 papers in the APS data as ranked by rescaled PageRank score $R(p)$ (asterisks mark the Milestone Letters).

Rank (p)	**Rank ($R(p)$)**	$p(\times 10^{-5})$	$R(p)$	Title	Year	Journal
1	1	43.32	29.96	Self-consistent equations including exchange and correlation effects (W. Kohn, L. Sham)	1965	*Phys. Rev.*
63	2	11.35	29.63	* Bose–Einstein condensation in a gas of sodium atoms (K. Davis et al.)	1995	*Phys. Rev. Lett.*
16	3	17.74	29.34	Self-organized criticality: an explanation of the $1/f$ noise (P. Bak, C. Tang, K. Wiesenfeld)	1987	*Phys. Rev. Lett.*
115	4	8.60	29.16	*Large mass hierarchy from a small extra dimension (L. Randall)	1999	*Phys. Rev. Lett.*
29	5	14.99	29.01	Pattern formation outside of equilibrium (M. Cross)	1993	*Rev. Mod. Phys.*
112	6	8.66	28.97	Statistical mechanics of complex networks (R. Albert, A.-L. Barabási)	2002	*Rev. Mod. Phys.*
181	7	7.11	28.95	Review of particle properties (K. Hagiwara et al)	2002	*Phys. Rev. D*
3	8	35.88	28.58	Inhomogeneous electron gas (P. Hohenberg)	1964	*Phys. Rev.*
99	9	9.35	28.58	Evidence of Bose–Einstein condensation in an atomic gas with attractive interactions (C. Bradley et al.)	1995	*Phys. Rev. Lett.*
59	10	11.65	28.11	Efficient pseudopotentials for plane-wave calculations (N. Troullier, J. Martins)	1991	*Phys. Rev. B*
53	11	12.11	27.88	*Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels (C. Bennett et al.)	1993	*Phys. Rev. Lett.*
281	12	5.99	27.85	*Negative refraction makes a perfect lens (J. Pendry)	2000	*Phys. Rev. Lett.*
216	13	6.59	27.59	Tev scale superstring and extra dimensions (G. Shiu, S.-H. Tye)	1998	*Phys. Rev. D*
17	14	17.54	27.47	Diffusion-limited aggregation, a kinetic critical phenomenon (T. Witten)	1981	*Phys. Rev. Lett.*
9	15	22.64	27.44	Self-interaction correction to density-functional approximations for many-electron systems (J. Perdew, A. Zunger)	1981	*Phys. Rev. B*

in their top-15, the detailed performance analysis described in the previous section is essential in order to fully understand the behavior of the two metrics.

4. Discussion

Motivated by the recent publication of the list of Milestone Letters by the Physical Review Letters editors, we performed an extensive cross-evaluation of different data-driven metrics of scientific impact of research papers with respect to their ability to identify papers of exceptional significance. We studied the network of citations between papers in the Physical Review corpus, which is recognized to be a comprehensive proxy for scientific research in physics (Radicchi & Castellano, 2011; Radicchi et al., 2009; Redner, 2005). The main assumption of our analysis is that although not all the most important papers in the Physical Review corpus are covered by the Milestone Letters list, a good paper-level metric is expected to rank the Milestone Letters as high as possible due to their outstanding significance. We find a clear performance gap between network-based metrics (p, $R(p)$, T) and local metrics based only on the number of received citations (c, $R(c)$). This finding suggests that the use of citation counts to rank scientific papers is sub-optimal; additional research will be needed to assess whether network-based article-level metrics can be used to construct author-level metrics more effective than the currently used metrics – such as the popular h-index introduced by Hirsch (2005) – that are only based on citation counts and neglect network centrality.

We have shown that the proposed rescaled PageRank $R(p)$ suppresses PageRank's well-known bias against recent papers much better than the CiteRank algorithm does. As a result, the proposed rescaled PageRank $R(p)$ provides a superior performance than PageRank and CiteRank in ranking recent and old milestone papers, respectively. There are still two possible ranking errors – false positives and false negatives – that have not been addressed in this manuscript. Young papers at the top of the ranking by the rescaled PageRank may be false positives because the citation spurt that they have experienced may stop which will eventually force them out of the ranking's top as well as out from the group of possibly highly significant papers. By contrast, the so-called “sleeping beauties” that receive a large part of citations long after they are published (Ke, Ferrara, Radicchi, & Flammini, 2015) are likely to be under-evaluated by the rescaled PageRank. Assessing the extent to which false positives and false negatives affect the ranking by rescaled PageRank, and by other relevant metrics as well, goes beyond the scope of our paper yet it constitutes a much needed step in future research. The analysis of larger datasets which include papers from diverse fields is another natural next step for future research. As different academic disciplines adopt different citation practices (Bornmann & Daniel, 2008), the rescaling procedure proposed in this paper may need to be extended to also remove possible ranking biases by academic field.

The assumptions behind our definition of time balance and the computation of the rescaled scores deserve attention as well. In agreement with Radicchi et al. (2008) and Radicchi and Castellano (2011), the definition of time balance of a ranking adopted in this article requires that the likelihood that a paper is ranked at the top by a time-balanced metric is independent of paper age. Our definition of ranking time balance is implicitly based on the assumption that the number of highly significant papers grows linearly with system size. While this assumption seems reasonable for the Physical Review corpus whose journals apply strict acceptance criteria for submitted papers, it might need to be reconsidered when analyzing larger datasets which include recently emerging high-acceptance journals – both mega-journals (Björk., 2015) and predatory journals (Xia et al., 2015). In other words, the exponential growth of the number of published papers (Redner, 2005; Sinatra, Deville, Szell, Wang, & Barabási, 2015; Wang, Song, & Barabási, 2013) does not necessarily correspond to an exponential growth of the number of highly significant papers. The issue is delicate (see Sarewitz, 2016 for a recent insight) and will need to be addressed in future research on bibliometric indicators.

An important general question remains open: which inherent properties of a network determine if PageRank-like methods will outperform local metrics or not? We conjecture that in citation networks, the observed success of network-based metrics in identifying highly significant papers might be related to the tendency of high-impact papers to cite other high-impact papers, as found by Bornmann, de Moya Anegón, and Leydesdorff (2010). Despite recent efforts (Fortunato, Boguñá, Flammini, & Menczer, 2008; Ghoshal & Barabási, 2011; Mariani et al., 2015; Medo, Mariani, Zeng, & Zhang, 2015), which network properties make the PageRank algorithm succeed or fail remains a largely unexplored problem which we will further investigate in future research.

Our work constitutes a particular instance of a general methodology – the comparison of the outcomes of quantitative variables with a ground-truth established by experts – which can be applied for metric evaluation in several kinds of systems, such as movies (Spitz & Horvát, 2014; Wasserman et al., 2015) or the network of scientific authors (Radicchi et al., 2009). In the domain of research evaluation, this methodology is particularly relevant since bibliometric indices are increasingly used in practice – often uncritically and in questionable ways (Hicks et al., 2015; Wilsdon, 2015) – and scholars from diverse field have produced a plethora of possible impact metrics (Van Noorden, 2010), especially those aimed at assessing researchers' productivity and impact. Motivated by the results obtained in this article, we encourage the creation of lists of groundbreaking papers also for other scientific domains, which can lead to a richer understanding and more accurate benchmarking of quantitative metrics for scientific significance. Our findings constitute a benchmark for article-level metrics of scientific significance, and can be used as a baseline to assess the performance of new indicators in future research.

From a practical point of view, improving the effectiveness of paper impact metrics has the potential to improve not only the current bibliometric practices, but also our ability to discover relevant papers in online platforms that collect academic papers and use automated methods to sort them. In this respect, our findings suggest that rescaled PageRank can be used as an operational tool to identify the most significant papers on a given topic. Suppose that a researcher enters a new research field and wants to study the most important works in that field. If we provide him/her with the top papers as ranked by PageRank, the researcher will only know the oldest papers and will not be informed about recent lines of research. On the other hand, by providing him/her with the top papers as ranked by rescaled PageRank, he/she will know both old significant papers and recent works that have attracted considerable attention, leading to a more complete overview of the field. To allow researchers to experience the benefits of a time-balanced ranking method, we developed an interactive Web platform which is available at the address http://www.sciencenow.info. In this platform, users can browse the rankings of the APS papers by $R(p)$ year by year, investigate the historical evolution of each paper's ranking position by $R(p)$, and check the ranking positions and the scores of each researcher's publications.

5. Conclusions

We presented a detailed analysis of the performance of different quantitative metrics with respect to their ability to identify the Milestone Letters selected by the Physical Review Letters editors. Our findings indicate that: (1) a direct rescaling of citation count and PageRank score is an effective way to suppress the temporal bias of these two metrics; (2) rescaled PageRank $R(p)$ is the best-performing metric overall, as it outperforms PageRank and CiteRank in identifying recent and old milestone papers, respectively, and it outperforms citation-based indicators for papers of every age. The presented results indicate that the combination of network centrality and time holds promise for improving some of the tools currently used to rank scientific publications, which could bring valuable benefits for quantitative research assessment and design of Web academic platforms.

Acknowledgements

We wish to thank Giulio Cimini, Matthieu Cristelli, Luciano Pietronero, Zhuo-Ming Ren, Ming-Sheng Shang, Andrea Tacchella, Giacomo Vaccario, Alexandre Vidmer and Andrea Zaccaria for inspiring discussions and useful suggestions. We are also grateful to the two anonymous referees for their insightful comments which helped us to improve the level of the discussion in some sections of the manuscript. This work was supported by the EU FET-Open Grant No. 611272 (project Growthcom). The authors declare that they have no competing financial interests.

Author contribution statement

Conceived and designed the analysis: Manuel Sebastian Mariani; Matúš Medo; Yi-Cheng Zhang
Collected the data: Manuel Sebastian Mariani; Matúš Medo
Contributed data or analysis tools: Manuel Sebastian Mariani; Matúš Medo
Performed the analysis: Manuel Sebastian Mariani
Wrote the paper: Manuel Sebastian Mariani; Matúš Medo

Appendix A. Average ranking position vs. average ranking ratio

We show here that the average ranking position of the MLs is extremely sensitive to the ranking position of the least-cited MLs, whereas the average ranking ratio is stable with respect to removal of the least-cited MLs. For simplicity, in this Appendix we consider the rankings computed on the whole dataset. In formulas, the average ranking position $r_{\bar{raw}}(s)$ of the MLs by metric s is defined as

$$r_{\bar{raw}}(s) = \frac{1}{M}\sum_{i \in \mathcal{M}} r_i(s), \tag{A.1}$$

where $r_i(s)$ denotes the ranking position of paper i by metric s normalized by the total number of papers: $r_i = 1/N$ and $r_i = 1$ correspond to the best and the worst paper in the ranking, respectively.

In Section 3.2, we mention that little-cited papers can bias the average ranking position of the target papers by a certain metric. To illustrate this point, consider first the following ideal example. Consider two target papers A and B. Paper A is ranked 10th by metric M_1 and 1000th by metric M_2, whereas paper B is ranked 20,000 by metric M_1 and 15,000 by metric M_2. The average ranking position for the set of papers $\{A, B\}$ is equal to 10,005 and to 8000 for metric M_1 and M_2, respectively. This means that according to average ranking position, metric M_2 outperforms metric M_1, despite having not been able to place any of the two papers in the top-100.

A qualitatively similar situation occurs also in the APS dataset, as the following example shows. The milestone letter "Element No. 102" [Phys. Rev. Lett. 1.1 (1958): 18] is cited only five times within the APS data. Its ranking position by $R(p)$ $(r(R(p)) = 0.22)$ is thus much larger than the MLs' average ranking position $r_{\bar{raw}}(R(p)) = 0.016$ by $R(p)$. Only few MLs are little cited – for instance, only four out of 87 MLs are not among the top-10% papers by citation count. To which extent do these little-cited papers affect $r_{\bar{raw}}$ for the different metrics? By denoting with $\bar{r'}_{raw}(R(p))$ the average computed on the subset of 83 MLs which does not include the four least-cited MLs, we obtain $\bar{r'}_{raw}(R(p)) = 0.009$, which is smaller than $\bar{r}_{raw}(R(p)) = 0.016$ by a factor around 1.8. The effect is even larger for citation count: we have $\bar{r'}_{raw}(c) = 0.009$ against the original value $\bar{r}_{raw}(c) = 0.020$ – the ratio between the two averages is larger than two.

By using the average ranking ratio, we only compare the ranking within the chosen set of metrics *for each individual paper* and, as a consequence, the average is stable with respect to removal of the least-cited MLs. This can be illustrated by again excluding the four least-cited MLs from the computation of $\bar{r}(R(p))$, and by comparing the corresponding values $\bar{r'}(R(p))$ of the average ranking ratio with the values computed over all the MLs. Among the five metrics, the largest variation

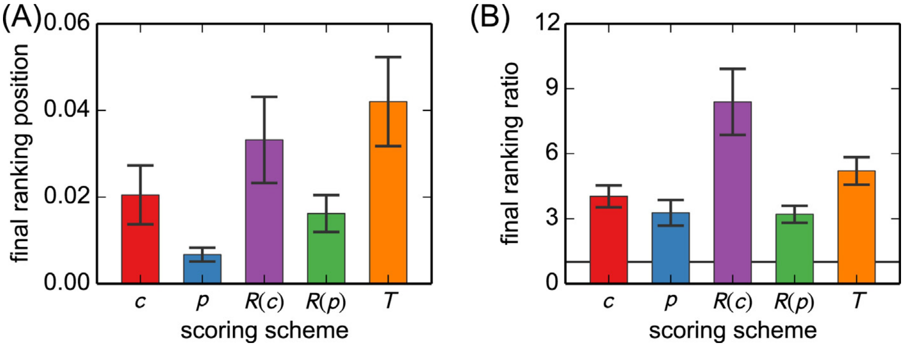

Fig. B.5. Values of the average ranking position $\bar{r_{raw}}$ (panel A) and of the average ranking ratio $\bar{r}$ (panel B) of the MLs for the five metrics computed on the whole dataset (1893–2009); the error bars represent the standard error of the mean.

is observed for PageRank, for which $\bar{r}'(p)/\bar{r}(p) = 1.03$ – i.e., the removal of the least-cited MLs has only a small effect on the average ranking ratios for the five metrics.

Appendix B. Assessing the metrics' performance on the whole dataset

Fig. B.5A shows the values of the average ranking position $\bar{r_{raw}}(s)$ for the five metrics computed on the whole dataset: according to $\bar{r_{raw}}(s)$, PageRank and rescaled PageRank outperform the other metrics.

While the average ranking position of the MLs is a simple quantity to evaluate the metrics, some MLs are relatively little cited and, as a result, their low ranking position can strongly bias the average ranking position. We refer to Appendix A for a detailed discussion of this issue. To solve this problem, we defined the ranking ratio in the main text. Fig. B.5B shows the measured values of the average ranking ratio $\bar{r}$ based on the rankings computed on the whole dataset. This simple measure would suggest that $R(p)$ and, to a lesser extent, p and c outperform $R(c)$ and CiteRank. Given the small gap between p and $R(p)$, one might be tempted to conclude that the rescaling procedure does not bring substantial benefits in the identification of significant papers. However, the rank analysis presented in Fig. B.5 includes the contribution of both old and recent MLs, whereas a close inspection reveals that the metrics perform in a drastically different way depending on the age of the target papers, as shown in Fig. 3 and discussed in Section 3.2.

This point can be also illustrated by using the rankings computed on the whole dataset. To show this, we divide the 87 MLs into three equally-sized groups of MLs according to their age. By considering only the oldest $M/3$ =29 MLs as target papers, we obtain $\bar{r}(p) = 1.1$ whereas $\bar{r}(R(p)) = 5.5$. By contrast, by considering only the $M/3$ most recent MLs as target papers, we obtain $\bar{r}(p) = 7.3$ whereas $\bar{r}(R(p)) = 1.7$. While this result shows a clear advantage of PageRank and rescaled PageRank for the oldest and for the most recent MLs, respectively, there exists a fundamental difference between the performance gaps observed for the oldest and the most recent MLs. The bias of PageRank towards old nodes (Fig. 1A) makes it indeed easier for the metric to find old significant papers. On the other hand, rescaled PageRank does not benefit from any bias in ranking the most recent MLs as the ranking by the metric is not biased by paper age (Fig. 1C). It is thus crucial to realize that when we compute the rankings on the whole dataset, the value of the average ranking ratio by the metrics depends on the age distribution of the important papers that we aim to identify. Were we using the rankings computed on the whole dataset for evaluation and were we only considering the oldest (most recent) 29 MLs as target papers, we would have concluded that PageRank (rescaled PageRank) is by far the best-performing metric. These observations demonstrate that an evaluation of the metrics based on the whole dataset is strongly biased by the age distribution of the target items and, for this reason, unreliable as a tool to assess metrics' performance.

Appendix C. Alternative rescaling equations

Eq. (3) forces the rescaled score $R_i(p)$ of a paper i to have mean value equal to zero and standard deviation equal to one, independently of its age (i.e., independently of i). Fig. 2C shows that this rescaling is sufficient to achieve a time-balanced ranking of the papers. We consider now a simple rescaling in the form $R_i^{(ratio)}(p) := p_i/\mu_i(p)$. While the mean value of this score is equal to one, one can show that its standard deviation is given by

$$\sigma\left[R_i^{(ratio)}(p)\right] = \sqrt{\mathrm{E}_i\left[(R_i^{(ratio)}(p))^2\right] - \mathrm{E}_i\left[R_i^{(ratio)}(p)\right]^2} = \sqrt{\frac{\mathrm{E}_i[p_i^2]}{\mu_i(p)^2} - 1} = \frac{\sigma_i(p)}{\mu_i(p)}, \tag{C.1}$$

where $E_i[\cdot]$ denotes the expectation value within the averaging window of paper i. Fig. C.6 shows that $\sigma(p)/\mu(p)$ strongly depends on node age in the APS dataset. As a result, the ranking by $R^{(ratio)}(p)$ is strong biased towards old nodes ($\sigma/\sigma_0 - 1$ =79.81 σ_{dev}).

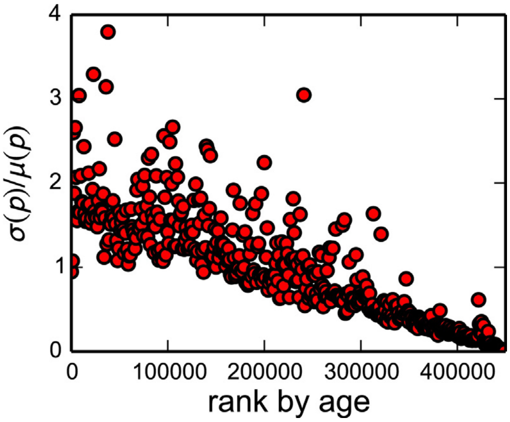

Fig. C.6. Dependence of $\sigma(p)/\mu(p)$ on paper age; the values of $\mu(p)$ and $\sigma(p)$ are calculated over the papers' averaging windows.

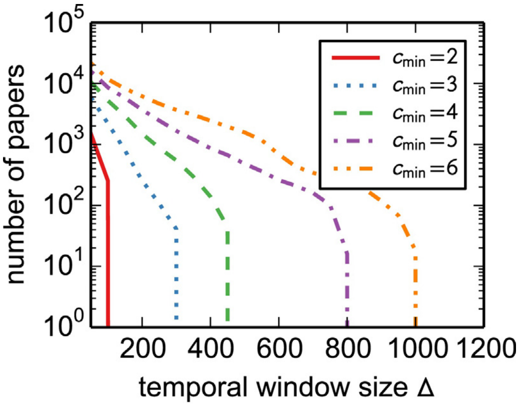

Fig. D.7. Number of papers whose averaging window contains less than five papers that received at least c_{min} citations as a function of Δ. For $\Delta \geq 1000$, each paper is compared with at least five papers cited at least five times.

We also considered a variant of our method where the rescaled scores are still computed with Eq. (3), but $\mu_i(p)$ and $\sigma_i(p)$ are computed over the papers published in the same year as paper *i*. The resulting rescaled score $R^{(year)}(p)$ produces a ranking that is much less in agreement with the hypothesis of unbiased ranking ($\sigma/\sigma_0 - 1 = 15.55\,\sigma_{dev}$) than the ranking by $R(p)$. For this reason, the definition of papers' averaging window adopted in the main text is based on number of publications and not on real time. However, $R^{(year)}(p)$ is still preferable to the original scores when the aim is to compare papers of different age. Also note that $R^{(year)}(p)$ might be preferable if one is interested in a ranking of the papers where each publication year is represented by the same number of papers, apart from statistical fluctuations.

Appendix D. Dependence of the properties of the rankings by *R*(*c*) and *R*(*p*) on the temporal window size Δ

As described in the main text, the rescaled scores $R_i(c)$ and $R_i(p)$ of a certain paper *i* are obtained by comparing its score with the scores of the nodes that belong to its "averaging windows" $j \in [i - \Delta_c/2, i + \Delta_c/2]$ and $j \in [i - \Delta_p/2, i + \Delta_p/2]$, respectively. To motivate the choice $\Delta_p = \Delta_c = 1000$ adopted in the main text, we start by observing that the size of the averaging window should be neither too large nor too small. A large window would include papers of significantly different age, which would turn out to be ineffective in removing the temporal biases of the metrics.[4] On the other hand, we want Δ_c and Δ_p to be sufficiently large to avoid that some papers are only compared with little-cited papers, which is likely to happen for a small window due to the skewed shape of the citation count distribution Medo et al. (2011).

To understand the possible drawbacks of a too small averaging window, we compute the number $N(c_{min})$ of papers whose averaging windows contain less than five papers that received at least c_{min} citations. The results are shown in Fig. D.7. For $\Delta \leq 800$, the averaging windows of a nonzero number of papers have less than five papers with at least five received citations. We restrict our choice to the range $\Delta \geq 1000$, for which no paper's average window has less than five papers cited at least five times.

[4] Note that the ranking by $R(p)$ is perfectly correlated with the ranking by *p* for $\Delta_p = N$.

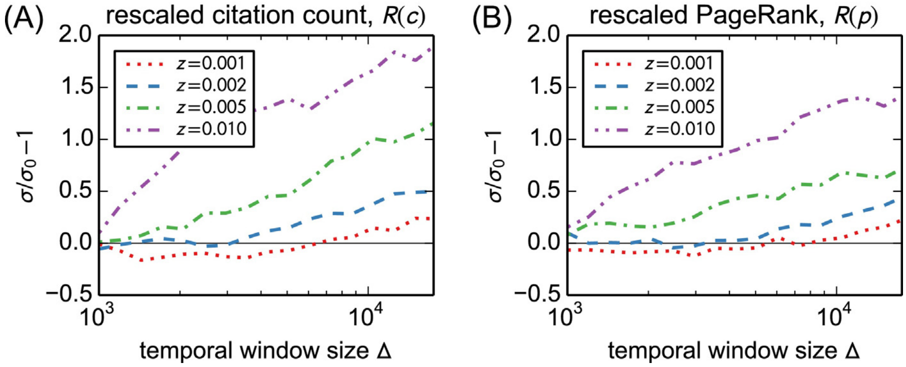

Fig. D.8. *Left panel*: Deviation $\sigma/\sigma_0 - 1$ for the ranking by rescaled citation count $R(c)$ as a function of Δ_c for different values of z. *Right panel*: Deviation $\sigma/\sigma_0 - 1$ for the ranking by rescaled PageRank score $R(p)$ as a function of Δ_p for different values of z. The horizontal black line marks the expected value $\sigma/\sigma_0 - 1$ =0 for an unbiased ranking.

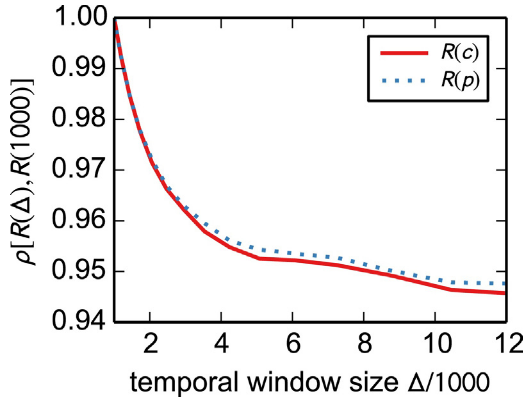

Fig. D.9. Spearman's ranking correlation between the rescaled score $R(\Delta)$ and the rescaled score $R(\Delta = 1000)$ used in the main text.

To evaluate the ability of the rescaling procedure to suppress the bias of the metrics, we estimate the deviation $\sigma/\sigma_0 - 1$ of the standard deviation ratio σ/σ_0 from the expected value (one) for an unbiased ranking (see the main text for details). Fig. D.8 reports the behavior of the deviation $\sigma/\sigma_0 - 1$ as a function of Δ_p and Δ_c for different selectivity values z. The upward trends of Fig. D.8 suggest that in order to reduce the ratio σ/σ_0, it is convenient to choose Δ_p and Δ_c as small as possible. Hence, the choice $\Delta_c = \Delta_p = 1000$ allows us to obtain an histogram close to the expected unbiased histogram – σ/σ_0 values are close to one for all the values of z represented in the figure – and, at the same time, to avoid that some nodes are only compared with little cited nodes, as discussed above for D.7.

An important observation is that the correlations between the rankings obtained with different values of Δ and the ranking obtained with $\Delta = 1000$ are close to one (Fig. D.9), which means that the rescaling procedure is robust against variation of the averaging window sizes Δ_c and Δ_p.

Appendix E. Dependence of CiteRank performance on its parameter τ

Fig. E.10 shows the dependence of the average ranking ratio $\bar{r}$ on paper age, for five different values of CiteRank parameter τ. The figure shows that the behavior of CiteRank's performance strongly depends on the choice of its parameter. When the parameter is small (panel A, $\tau = 1$ year), CiteRank performance is optimal (lowest average ranking ratio) for very recent papers, and gradually worsens with paper age. As τ increases (moving from panel A to E), the minimum point of CiteRank's average ranking ratio gradually shifts toward older nodes. When τ is sufficiently large (panel E, $\tau = 16$ years), CiteRank behavior is qualitatively similar to that of PageRank, and its performance gradually improves with paper age – this is indeed consistent with the fact that $T \to p$ in the limit $\tau \to \infty$.

Appendix F. Dependence of ranking ratio and identification rate on paper age

To assess the ranking of each Milestone Letter t years after its publication, we compute the rankings each $\Delta t = 183$ days (results for different choices of Δt are qualitatively similar). At each computation time $t^{(c)}$, only the $N(t^{(c)})$ papers (with their links) published before time $t^{(c)}$ are considered for the scores' and rankings' computation, and each ML contributes to the

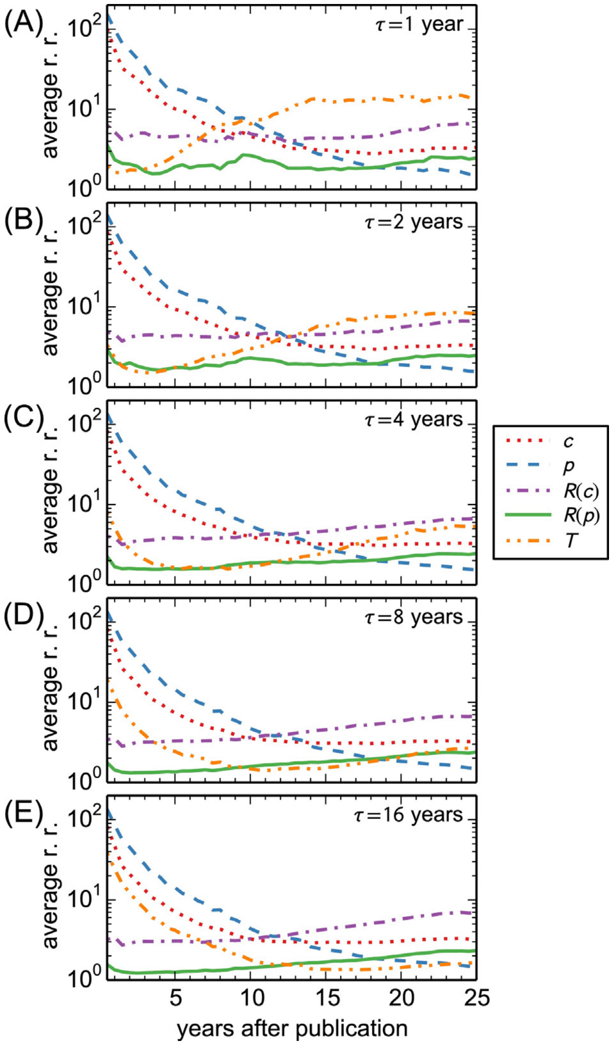

Fig. E.10. Dependence of the average ranking ratio $\bar{r}$ on paper age, for five different values of CiteRank parameter τ.

ranking ratio $\bar{r}(s, t)$ corresponding to its age t at time $t^{(c)}$. This procedure allows us to save computational time with respect to computing the rankings of each ML *exactly* t years after its publication, because it requires fewer ranking computations.

In formulas, the average ranking ratio $\bar{r}(s, t = k\,\Delta t)$ for t-years old papers is defined as

$$\bar{r}(s, t = k\,\Delta t) = \frac{1}{M(t)} \sum_{t^{(c)}} \sum_{i \in \mathcal{M}} \delta\left(\lfloor (t^{(c)} - t_i)/\Delta t \rfloor, k\right) \times \frac{r(s, i; t^{(c)})}{\min_{s'}\{r(s', i; t^{(c)})\}}, \tag{F.1}$$

where we used k = 0.5, 1, 1.5, 2, . . . for Fig. 3B; in the equation above, $r(s, i; t^{(c)})$ denotes the ranking position of ML i at time $t^{(c)}$ according to metric s, $M(t)$ denotes the number of MLs that are at least t years old at the end of the dataset, $\lfloor x \rfloor$ denotes the largest integer smaller than or equal to x, $\delta(x, y)$ denotes the Kronecker delta function of x and y. Hence, at each computation time $t^{(c)}$, each ML i published before time $t^{(c)}$ gives a contribution $\hat{r}(s, i; t^{(c)})$ to the average ranking ratio $\bar{r}(s, t = k\,\Delta t)$ for papers of age $t^{(c)} - t_i$. Similarly, the identification rate $f_x(t)$ is computed as

$$f_x(s, k\,\Delta t) = \frac{1}{M(t)} \sum_{t^{(c)}} \sum_{i \in \mathcal{M}} \delta\left(\lfloor (t^{(c)} - t_i)/\Delta t \rfloor, k\right) \times \chi(r(s, i; t^{(c)}) \leq x), \tag{F.2}$$

where $\chi(r(s, i; t^{(c)}) \leq x)$ is equal to one if paper i is among the top $x\,N(t^{(c)})$ papers in the ranking by metric s at time $t^{(c)}$, equal to zero otherwise.

To define the normalized identification rate (NIR) of a metric, at each computation time $t^{(c)}$ we divide the $N(t^{(c)})$ papers into 40 groups according to their age, analogously to what we did in Section 3.1 to evaluate the temporal balance of the metrics. The NIR of metric s is then defined as

$$\tilde{f}_x(s, k\,\Delta t) = \frac{1}{M(t)} \sum_{t^{(c)}} \sum_{i \in \mathcal{M}} \delta\left(\lfloor (t^{(c)} - t_i)/\Delta t \rfloor, k\right) \times \chi(r(s, i; t^{(c)}) \leq x)\, y(n(s, i; t^{(c)})), \tag{F.3}$$

where $y(n(s, i;t^{(c)}))$ is a decreasing function of the fraction $n(s, i;t^{(c)})$ of nodes that belong to the same age group of node i and are ranked among the top $x\,N(t^{(c)})$ by metric s. Denoting by $n_0(i;t^{(c)}) = 1/40$ the expected value of $n(\cdot, i;t^{(c)})$ for an unbiased ranking, we set $y(n(s, i; t^{(c)})) = (n(s, i; t^{(c)})/n_0(i; t^{(c)}))^{-1}$ if $n(s, i;t^{(c)}) > n_0(i;t^{(c)})$ (i.e., if the metric tends to favor papers that belong to the same age group as paper i), whereas $y(n(s, i;t^{(c)})) = 1$ if $n(s, i;t^{(c)}) \leq n_0(i;t^{(c)})$. According to Eq. (F.3), if the identified ML belongs to an age group which is over-represented in top $x\,N(t^{(c)})$ by the factor of four, it only counts as 1/4 in the normalized identification rate.

Appendix G. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.joi.2016.10.005.

References

Berkhin, P. (2005). A survey on pagerank computing. *Internet Mathematics*, *2*(1), 73–120.

Björk., B.-C. (2015). Have the "mega-journals" reached the limits to growth? *PeerJ*, *3*, e981.

Bollen, J., Rodriquez, M. A., & Van de Sompel, H. (2006). Journal status. *Scientometrics*, *69*(3), 669–687.

Bornmann, L., & Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80.

Bornmann, L., de Moya Anegón, F., & Leydesdorff, L. (2010). Do scientific advancements lean on the shoulders of giants? A bibliometric investigation of the ortega hypothesis. *PLoS ONE*, *5*(10), e13327.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, *30*(1), 107–117.

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

Cimini, G., Gabrielli, A., & Labini, F. S. (2014). The scientific competitiveness of nations. *PLOS ONE*, *9*(12), e113470.

Crespo, J. A., Ortuño-Ortín, I., & Ruiz-Castillo, J. (2012). The citation merit of scientific publications. *PloS one*, *7*(11), e49156.

Dunaiski, M., & Visser, W. (2012). Comparing paper ranking algorithms. In *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference* (pp. 21–30). ACM.

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

Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics*, *69*(1), 131–152.

Ermann, L., Frahm, K. M., & Shepelyansky, D. L. (2015). Google matrix analysis of directed networks. *Reviews of Modern Physics*, *87*(4), 1261.

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

Fortunato, S., Boguñá, M., Flammini, A., & Menczer, F. (2008). Approximating pagerank from in-degree. In *Algorithms and models for the web-graph* (pp. 59–71). Springer.

Franceschet, M. (2011). Pagerank: Standing on the shoulders of giants. *Communications of the ACM*, *54*(6), 92–101.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Ghoshal, G., & Barabási, A.-L. (2011). Ranking stability and super-stable nodes in complex networks. *Nature Communications*, *2*, 394.

Gleich, D. F. (2015). Pagerank beyond the web. *SIAM Review*, *57*(3), 321–363.

González-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals' scientific prestige: The SJR indicator. *Journal of Informetrics*, *4*(3), 379–391.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*, 429–431.

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

Kaur, J., Radicchi, F., & Menczer, F. (2013). Universality of scholarly impact metrics. *Journal of Informetrics*, *7*(4), 924–932.

Kaur, J., Ferrara, E., Menczer, F., Flammini, A., & Radicchi, F. (2015). Quality versus quantity in scientific impact. *Journal of Informetrics*, *9*(4), 800–808.

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431.

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311–316.

Kinney, A. L. (2007). National scientific facilities and their science impact on nonbiomedical research. *Proceedings of the National Academy of Sciences*, *104*(46), 17943–17947.

Kreyszig, E. (2010). *Advanced engineering mathematics*. John Wiley & Sons.

Lawrence, P. A. (2008). Lost in publication: How measurement harms science. *Ethics in Science and Environmental Politics*, *8*, 9–11.

Liebowitz, S. J., & Palmer, J. P. (1984). Assessing the relative impacts of economics journals. *Journal of Economic Literature*, *22*(1), 77–88.

Lü, L., Medo, M., Yeung, C. H., Zhang, Y.-C., Zhang, Z.-K., & Zhou, T. (2012). Recommender systems. *Physics Reports*, *519*(1), 1–49.

Mariani, M. S., Medo, M., & Zhang, Y.-C. (2015). Ranking nodes in growing networks: When pagerank fails. *Scientific Reports*, *5*

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

Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters*, *107*(23), 238701.

Medo, M., Mariani, M. S., Zeng, A., & Zhang, Y.-C. (2015). *Identification and modeling of discoverers in online social systems.*, arXiv preprint arXiv:1509.01477.

Molinari, J.-F., & Molinari, A. (2008). A new methodology for ranking scientific institutions. *Scientometrics*, *75*(1), 163–174.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: Computer Horizons.

Newman, M. (2010). *Networks: An introduction*. Oxford University Press.

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

Newman, M. E. J. (2014). Prediction of highly cited papers. *Europhys. Lett.*, *105*(2), 28002.

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

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

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

Radicchi, F., & Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E*, *83*(4), 046116.

Radicchi, F., & Castellano, C. (2012a]). A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. *PLoS ONE*, *7*(3), e33833.

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

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

Radicchi, F., Fortunato, S., Markines, B., & Vespignani, A. (2009). Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, *80*(5), 056103.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*, 49.

Sarewitz, D. (2016). The pressure to publish pushes down quality. *Nature*, *533*(7602), 147–147.

Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A.-L. (2015). A century of physics. *Nature Physics*, *11*(10), 791–796.

Spitz, A., & Horvát, E.-Á. (2014). Measuring long-term impact based on network centrality: Unraveling cinematic citations. *PLoS One*, *9*(10), e108857.

Van Noorden, R. (2010). Metrics: A profusion of measures. *Nature*, *465*(7300), 864–866.

Van Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, *62*(1), 133–143.

Walker, D., Xie, H., Yan, K.-K., & Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, *2007*(06), P06010.

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

Waltman, L., & Yan, E. (2014). Pagerank-related methods for analyzing citation networks. In *Measuring scholarly impact*. pp. 83–100. Springer.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.

Wasserman, M., Zeng, X. H. T., & Amaral, L. A. N. (2015). Cross-evaluation of metrics to estimate the significance of creative works. *Proceedings of the National Academy of Sciences*, *112*(5), 1281–1286.

Weingart, P. (2005). Impact of bibliometrics upon the science system: Inadvertent consequences? *Scientometrics*, *62*(1), 117–131.

Werner, R. (2015). The focus on bibliometrics makes papers less useful. *Nature*, *517*(7534), 245.

Wilsdon, J. (2015). We need a measured approach to metrics. *Nature*, *523*(7559), 129–129.

Xia, J., Harmon, J. L., Connolly, K. G., Donnelly, R. M., Anderson, M. R., & Howard, H. A. (2015). Who publishes in “predatory” journals? *Journal of the Association for Information Science and Technology*, *66*(7), 1406–1417.

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

Yao, L., Wei, T., Zeng, A., Fan, Y., & Di, Z. (2014). Ranking scientific publications: The effect of nonlinearity. *Scientific Reports*, *4*, 6663.

Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2015). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, 1–12.