Published in "Journal of Informetrics 11(3): 766–782, 2017"
which should be cited to refer to this work.

Quantifying and suppressing ranking bias in a large citation network

Giacomo Vaccario[a], Matúš Medo[b,c,d], Nicolas Wider[a], Manuel Sebastian Mariani[d,e,*]

[a] *Chair of Systems Design, ETH Zurich, 8092 Zurich, Switzerland*
[b] *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, PR China*
[c] *Department of Radiation Oncology, Inselspital, Bern University Hospital and University of Bern, 3010 Bern, Switzerland*
[d] *Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland*
[e] *Guangdong Province Key Laboratory of Popular High Performance Computers, College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, PR China*

It is widely recognized that citation counts for papers from different fields cannot be directly compared because different scientific fields adopt different citation practices. Citation counts are also strongly biased by paper age since older papers had more time to attract citations. Various procedures aim at suppressing these biases and give rise to new normalized indicators, such as the relative citation count. We use a large citation dataset from Microsoft Academic Graph and a new statistical framework based on the Mahalanobis distance to show that the rankings by well known indicators, including the relative citation count and Google's PageRank score, are significantly biased by paper field and age. Our statistical framework to assess ranking bias allows us to exactly quantify the contributions of each individual field to the overall bias of a given ranking. We propose a general normalization procedure motivated by the *z*-score which produces much less biased rankings when applied to citation count and PageRank score.

Keywords:
Impact indicators
Ranking
Network analysis
Field bias
Field normalization

1. Introduction

Paper citation count itself and various quantities derived from it are used as influential indicators of research impact (Garfield, 2006; Hirsch, 2005). At the same time, it is well known that the cumulative number of citations received by academic publications strongly depends on paper age and field (Schubert & Braun, 1986; Vinkler, 1986). Old papers have had more time to acquire citations than recent ones, and their advantage is further enhanced by the preferential attachment mechanism (de Solla Price, 1976; Newman, 2009). While heterogeneous paper fitness and paper aging possibly attenuate the advantage of old nodes (Medo, Cimini, & Gualdi, 2011; Wang, Song, & Barabási, 2013), empirical evidence typically shows that citation count is still biased toward old nodes (see Mariani, Medo, & Zhang, 2016; Newman, 2009; Radicchi & Castellano, 2011, among others). In addition, different academic fields adopt very different citation practices (see Bornmann & Daniel, 2008 for a review on the topic), which results in a strong dependence of the mean number of citations on academic field, as shown in several works (Bornmann & Daniel, 2009; Lundberg, 2007; Radicchi, Fortunato, & Castellano, 2008, among others).

* Corresponding author at: Department of Physics, University of Fribourg, 1700 Fribourg, Switzerland.
E-mail address: manuel.mariani@unifr.ch (M.S. Mariani).

A natural question arises: how can we "fairly" use citation-based indicators to compare papers from different fields and of different age? The problem of comparing papers from different fields is usually referred to as the *field-normalization* problem. Several approaches to address this question have been proposed in the literature (see Waltman, 2016 for a recent review). A particularly simple approach is to divide each paper's citation count by the mean number of citations for papers of the same field published in the same year. The results by Radicchi et al. (2008) suggested that this indicator, called *relative citation count*, produces a ranking that is statistically consistent with the hypothesis of a ranking that is not biased by field and age. This finding has been challenged by subsequent works by Albarrán, Crespo, Ortu no, and Ruiz-Castillo (2011) and Waltman, van Eck, and van Raan (2012), which leaves the debate on age- and field-normalization procedures still open.

In this article, we analyze a large dataset from Microsoft Academic Graph (Sinha et al., 2015) to show that existing indicators of impact, including the relative citation count, fail to produce rankings that are not biased by age and field. To simultaneously assess these biases, we present a new procedure based on the Mahalanobis distance (Mahalanobis, 1936). This permits to compare the ranking by a given indicator with those obtained with a simulated unbiased sampling, and hence to quantify the overall ranking bias. An analytic result derived in this paper allows us to assess the contribution of each field to the overall ranking bias. It is worth noticing that while we focus on the biases by age and field, our bias assessment procedure can be easily extended to detect any other kind of information bias.

We also present the first systematic study of the possible bias by field of the PageRank score (Brin & Page, 1998) and of its age-rescaled version introduced by Mariani et al. (2016) at article-level. The motivation to analyze these network-based indicators comes from the finding that they outperform other metrics in identifying expert-selected milestone papers (Mariani et al., 2016). However, the application of PageRank and its variants to academic citation networks focused on datasets composed of papers from a single field (Chen, Xie, Maslov, & Redner, 2007; Mariani et al., 2016; Mariani, Medo, & Zhang, 2015; Walker, Xie, Yan, & Maslov, 2007; Yao, Wei, Zeng, Fan, & Di, 2014; Zhou, Zeng, Fan, & Di, 2016). While the possible bias by scientific field of eigenvector-based algorithms has been explored by Waltman and van Eck (2010) at journal-level, the PageRank score's possible bias by academic field at article-level is (to our best knowledge) still unexplored and we are the first authors to address it.

We introduce two novel indicators of impact motivated by the z-score: age- and field-rescaled citation count $R^{AF}(c)$ and age- and field-rescaled PageRank $R^{AF}(p)$. We find that the novel indicators produce paper rankings that are much less biased by age and field than the rankings produced by the other analyzed indicators. Nevertheless, also the Mahalanobis distance observed for the new indicators is not statistically consistent with ones obtained for a simulated unbiased process. This indicates that the problem of achieving an ideal unbiased ranking of the publications remains open.

The rest of our article is organized as follows: Section 2 describes the analyzed dataset of publications obtained from the Microsoft Academic Graph. Section 3 presents existing paper-level impact indicators and reports their bias by scientific field. In Section 4, we introduce a rescaling procedure for citation count and PageRank scores motivated by the z-score. In Section 5, we introduce a general procedure to test for any kind of ranking bias, and present its application to assess the field and age bias of the rankings by the indicators studied here. In Section 6, we conclude by discussing possible limitations of our analysis and future research directions.

2. Data

We analyze a bibliographic dataset which was provided for the KDD Cup 2016.[1] This data is a dump of the *Microsoft Academic Graph* (MAG) and contains more than 126 millions of publications and more than 467 millions citations (Sinha et al., 2015). Each publication is also endowed with various properties such as unique ID, publication date, title and journal ID. We pre-processed the data (details are provided in Appendix A) to remove from the analysis papers with incomplete information, ending up with $N = 18\,193\,082$ unique publications and $E = 109\,719\,182$ citations.

The MAG has a field classification at paper level (Sinha et al., 2015). In the KDD cup dump, there are 19 main fields and numerous subfields up to 3 hierarchical levels of subsubfields. However, all the different subfields can belong to several main fields, meaning that each publication can belong to more than one main field. We use here the field classification at the highest hierarchical level, i.e., we only consider the 19 main fields. When calculating the citation count and PageRank score of papers (see Section 3), we consider the publications that belong to more than one field only once. In this way, we do not modify the number of citations that each paper receives and gives, and we do not change the topology of the network on which the PageRank scores are calculated. On the other hand, in agreement with Waltman et al. (2012), to compute the fields' size (see Table A.2 in Appendix A) and the field-rescaled metrics (see Sections 3 and 4), each publication can be considered multiple times in the analysis, once for each field the publication belongs to. In this way, each field is represented by all its publications even if some of these are shared with other fields.

Before moving to the next Sections, we devote our attention to two main assumptions of our analysis. First, we assume that the Microsoft Academic data provide a representative sample of the population of publications and of their citations. This assumption is motivated by the findings of independent analyses of the Microsoft Academic dataset (Harzing & Alakangas, 2013; Hug & Braendle, 2017) that have shown that its coverage is comparable to other popular academic databases, such as Scopus and Web of Science.

[1] https://kddcup2016.azurewebsites.net/Data

Second, to quantify the bias by field of impact indicators, we assume that the fields are given by the Microsoft Academic's field classification scheme at its highest hierarchical level. In the literature, there is no general agreement on which field classification scheme should be used to classify papers and there is an entire stream of works investigating issues related to this (Adams, Gurney, & Jackson, 2008; Colliander & Ahlgren, 2011; Radicchi & Castellano, 2012a; Sirtes, 2012; Zitt, Ramanana-Rahary, & Bassecoulard, 2005). In particular, the choice of a suitable aggregation level has been shown to be delicate: by considering the most aggregate fields, heterogeneities in the subfields' citation patterns might be hidden (Radicchi & Castellano, 2011; van Leeuwen & Medina, 2012) – this effect has been shown to be magnified when iterative ranking algorithms are used instead of citation count (Waltman, Yan, & van Eck, 2011). On the other hand, increasing the resolution of the field classification may lead to largely-overlapping fields or to hardly interpretable fields. For example, Hug, Ochsner, and Brandle (2017) show that the MAG fields at the second highest level are too detailed and, for this reason, the authors suggest that they should not be used for field-normalization purposes. We leave to future research the important study of how different classification schemes impact the biases of rankings and how our results generalize to other data sets.

3. Shortcomings of existing metrics

We now define the four existing metrics analyzed in this work: citation count c, relative citation count c^f, PageRank score p, age-rescaled PageRank score $R^A(p)$ (Section 3.1). Furthermore, we show that these metrics are severely biased by scientific field (Section 3.2).

3.1. Definition of existing metrics

Citation count, c. The citation count c_i of node i is simply the number of citations received by paper i. In terms of the citation network's adjacency matrix $\mathbf{A}$ (in a directed network, $A_{ij} = 1$ if node j points to node i, $A_{ij} = 0$ otherwise), we can express the citation count as $c_i = \sum_j A_{ij}$.

Relative citation count, c^f. To overcome citation count's bias by paper age and academic field, Radicchi et al. (2008) defined the *relative citation count* c_i^f of paper i as $c_i^f := c_i/\mu_i^Y(c)$, where $\mu_i^Y(c)$ denotes the mean citation count for papers published in the same field and year as paper i. Throughout this paper, we always refer to the 19 main fields provided in the MAG dataset.

PageRank, p. Citation count and metrics built on it share an important limitation: the citations a paper receives are all counted the same, regardless of the importance of the citing paper. A possible way to overcome this limitation – recognized already in the 70s in the scientometrics community (Pinski & Narin, 1976) – is to take into account the whole structure of the paper-paper citation network. In this spirit, eigenvector-based metrics take as input the citation network's adjacency matrix $\mathbf{A}$. This class of metrics have been applied in various research domains including scientometrics (Bergstrom, West, & Wiseman, 2008; Pinski & Narin, 1976), Web information retrieval (Brin & Page, 1998; Kleinberg, 1999), social science (Bonacich, 1987; Katz, 1953) – see (Ermann, Frahm, & Shepelyansky, 2015; Franceschet, 2011; Gleich, 2015) for a review. Among these metrics, we focus on Google's PageRank score (Brin & Page, 1998). This was originally devised to rank webpages in the World Wide Web and has attracted considerable interest of the scientometrics community. The rationale behind its application to citation networks is that citations coming from influential papers should count more than citations from obscure articles.

The PageRank scores of papers are contained in a vector $\mathbf{p}$ defined by the following equation

$$\mathbf{p} = \alpha \mathbf{P}\,\mathbf{p} + (1-\alpha)\mathbf{v}, \tag{1}$$

where α is a parameter of the algorithm (called damping factor), $\mathbf{P}$ is the random-walk transition matrix with elements $P_{ij} = A_{ij}/k_j^{out}$, $k_j^{out} = \sum_l A_{lj}$ is the number of references in paper j, and $\mathbf{v}$ is a uniform teleportation vector with elements $v_i = 1/N$ for all papers i. Eq. (1) can be interpreted as the stationary equation of a stochastic process on the citation network. In this process, a random walker is placed on each paper and he/she either follows a citation edge with probability α, or jumps to a randomly chosen paper with probability $1-\alpha$. When the number of walkers on each paper reaches a stationary value, we obtain the PageRank score of a paper i by calculating the fraction of walker on this paper. There is no universal criterion to choose the value of the damping factor α. In agreement with Chen et al. (2007), we set $\alpha = 0.5$ which corresponds to a random walker covering paths of length two before teleporting to a random node, as opposed to paths of length close to seven expected with the often used value $\alpha = 0.85$. Chen et al. (2007) argue that the choice $\alpha = 0.5$ better reflects the actual surfing behavior of researchers than $\alpha = 0.85$.

PageRank is based on a static, time-aggregated perspective of the considered network. In general, such perspective has been shown to be limiting for the analysis of evolving networks (Mariani et al., 2015; Scholtes, Wider, Pfitzner, Garas, Tessone, & Schweitzer, 2014b, Wider, et al., 2014). While the resulting metric's bias towards old papers has already been studied in the literature (Chen et al., 2007; Mariani et al., 2015, 2016; Maslov & Redner, 2008), its possible bias by academic field is still unexplored and we address it in Section 3.2.

Age-rescaled PageRank, $R^A(p)$. To suppress the age bias of PageRank, Mariani et al. (2016) proposed to rescale the PageRank score by comparing each paper's score with the scores of papers of similar age. Assuming that the papers are ordered by

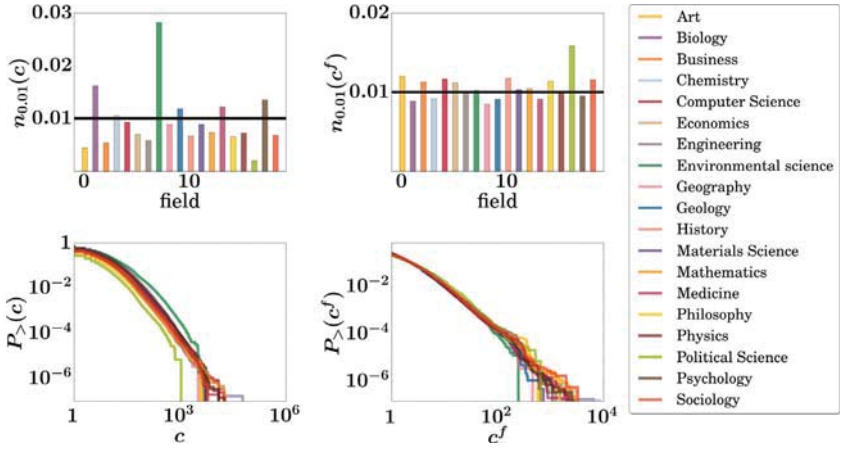

Fig. 1. Field bias of the analyzed citation-based metrics. Top panels show histograms of the fraction of top-1% publications for each field in the ranking by (left to right) citation count and relative citation count. The black horizontal line is at 0.01, i.e. the expected value. Bottom panels show for each field the complementary cumulative distributions for citation count (left) and relative citation count (right).

older to younger, one computes the mean value $\mu_i^A(p)$ and the standard deviation $\sigma_i^A(p)$ of PageRank scores over Δ_p papers around i, i.e. $j \in [i - \Delta_p/2, i + \Delta_p/2]$. Consequently, the rescaled PageRank score $R_i^A(p)$ of paper i is defined as

$$R_i^A(p) = \frac{p_i - \mu_i^A(p)}{\sigma_i^A(p)}. \tag{2}$$

Mariani et al. (2016) applied rescaled PageRank to the network of physics papers published by the American Physical Society journals to show that the resulting ranking is not biased by paper age and, as a result, it allows us to identify seminal publication much earlier than rankings by metrics that are biased against recent papers. In the following, we set $\Delta_p = 1000$ as in Mariani et al. (2016).

3.2. Field bias of the existing metrics

After having described a set of existing metrics, we now apply them to the MAG dataset to show that the rankings that they produce are biased by scientific field. For a ranking that is not biased by scientific field, the number of top-ranked publications from each field should be proportional to the total number of publications from that field. In other words, for an unbiased ranking, we expect

$$\mu_f = \frac{z}{100} K_f \tag{3}$$

papers from field f among the top z% papers in the ranking, where K_f is the total number of publications from field f (Radicchi et al., 2008). In the following, we denote by $k_f^{(m)}$ the number of publications from field f in the top-1% of the ranking by metric m. We restrict our analysis to z% =1%; results for other values of z are available upon request from the authors.

In the top panels of Fig. 1, we illustrate the field bias of citation count, c, and relative citation count, c^f. The presence of strong biases is evident for both metrics because there are fields whose ratio $k_f^{(m)}/K_f$ is far away from the expected value 0.01. In particular, Environmental Science is extremely over-represented in the top of the ranking by citation count. We argue that this bias comes from the fact that publications from this field have a mean citation count almost twice as big compared to publications belonging to other fields (see Table A.2). For relative citation count, we find a better agreement with what we would expect from an unbiased indicator. However, relatively large deviations are still evident, especially for the field of Political Science.

In the bottom panels of Fig. 1, we report the distributions of c and c^f for each field. These panels show that the bias by field is not limited to the top 1% papers in the ranking, but it arises from systematic differences between the score distributions across different fields. For example, when looking at the distribution of c, papers in the field of Political Science have consistently smaller probability to have more than one citation compared to other fields. For a detailed discussion about the bias of the ranking by c^f, we refer to Appendix C.

Fig. 2 reports the same analysis for PageRank scores, p, and age-rescaled PageRank scores, $R^A(p)$. This figure provides the first study of the dependence of PageRank score on academic field. The top panels of Fig. 2 show that the top positions of both rankings are biased by field, and both rankings overestimate the impact of publications in the field of Environmental Science. Again, we argue that this happens because the mean indegree of publications from Environmental Science is approximately twice as big compared to publications that belong to other fields (see Table A.2). From the bottom-left panel of Fig. 2, we find that the full distribution of scores of Page Rank have a similar shape, but different broadness. These differences are slightly

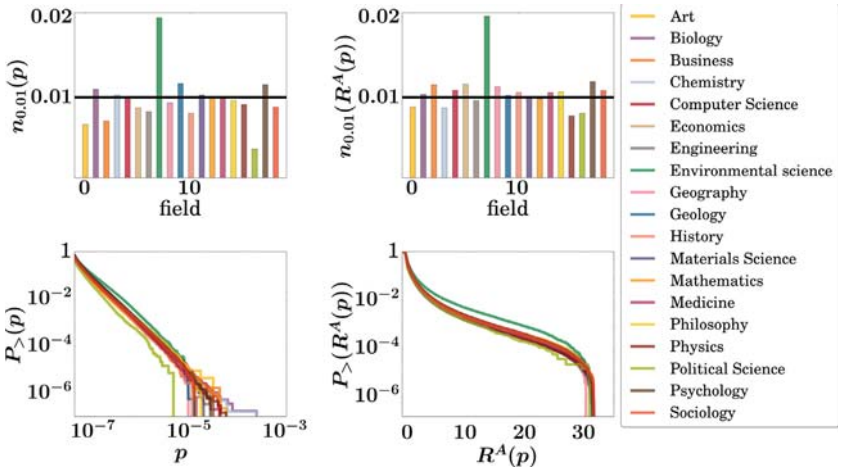

Fig. 2. Field bias of the analyzed measures based on PageRank. Top panels show histograms of the fraction of top-1% publications for each field in the ranking by (left to right): PageRank and age-rescaled PageRank. The black horizontal line is at 0.01, i.e. the expected value. Bottom panels show for each field the complementary cumulative distributions for PageRank (left) and age-rescaled PageRank (right).

smaller for the age-rescaled PageRank, with the exception of the field of Environmental Science (see bottom right panel of Fig. 2).

4. Defining new age- and field-normalized metrics

In this section, we introduce two novel indicators of paper impact: the age- and field-rescaled citation count, $R^{AF}(c)$, and the age- and field-rescaled PageRank, $R^{AF}(p)$. The two indicators, $R^{AF}(c)$ and $R^{AF}(p)$, are obtained from citation count c and PageRank score p, respectively, through a rescaling procedure. This procedure is based on the z-score and is aimed at suppressing age and field bias. The idea of using the z-score is not new in scientometrics (Bornmann & Daniel, 2009; Lundberg, 2007; Mariani et al., 2016; McAllister, Narin, & Corrigan, 1983; Newman, 2009; Zhang, Cheng, & Liu, 2014); our new indicators can be considered as variants of the indicator based on the z-score studied by Zhang et al. (2014) and their main difference is explained below.

4.1. Age- and field-rescaled citation count, $R^{AF}(c)$

To calculate the age- and field-rescaled citation count $R_i^{AF}(c)$ of a paper i belonging to a field f, we first compute the mean $\mu_i^{AF}(c)$ and the standard deviation $\sigma_i^{AF}(c)$ of the citation count of papers of the *same field* and of *similar age* as paper i. In particular, $\mu_i^{AF}(c)$ and $\sigma_i^{AF}(c)$ are computed over the papers that belong to the same field f as paper i and that are among the Δ_c closest papers to i as measured by the distance $|i-j|$ between their rank by age. Then, the age- and field-rescaled citation count score $R_i^{AF}(c)$ is defined as

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

The averaging window size Δ_c is a parameter of the method, which we set to $\Delta_c = 1000$.

Differently from Zhang et al. (2014), for the computation of the z-score, we use temporal windows with the same number of publications, which in general corresponds to real-time intervals of different duration. This choice is supported by recent findings (Parolo et al., 2015) that indicate that in citation networks, time is better defined by number of publications than by real time. Furthermore, rescaled metrics based on the z-score with fixed temporal-window duration have already been shown to underperform with respect to the relative citation count c^f in the task of producing an unbiased ranking (Zhang et al., 2014). Our analysis (not shown) confirms that when the temporal windows are of a fixed temporal length, the bias removal is inferior to that achieved with temporal windows containing a fixed number of publications. For these reasons, we do not include metrics based on z-score with fixed temporal-window duration in our analysis.

Differently from the relative citation count c^f, $R^{AF}(c)$ is expected to have not only uniform mean value across different publication dates and fields, but also uniform standard deviation. This should lead to a more balanced ranking of the papers. We show in the following that this is indeed the case.

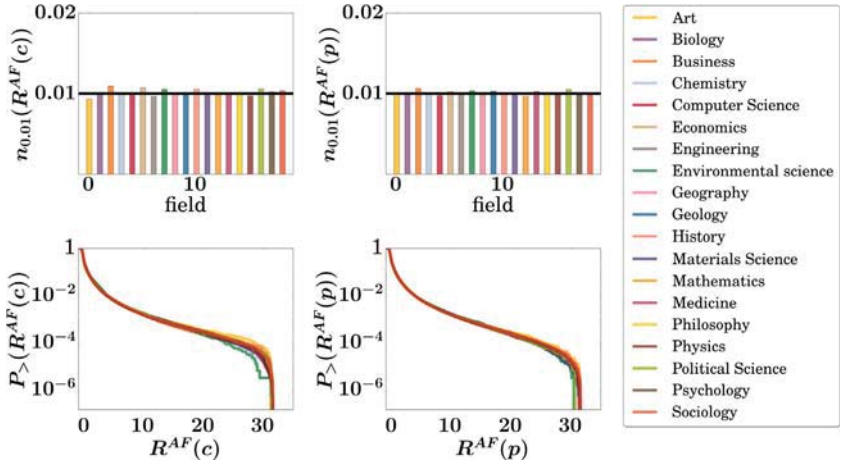

Fig. 3. Field balance of the analyzed citation-count and PageRank-based metrics. Top panels show histograms of the number of top-1% publications for each field in the ranking by (left to right): age- and field-rescaled citation count, and age- and field-rescaled PageRank. The black horizontal line is at 0.01, i.e. the expected value. Bottom panels show for each field the complementary cumulative distributions for age- and field-rescaled citation count (left), and age- and field-rescaled PageRank (right).

4.2. Age- and field-rescaled PageRank, $R^{AF}(p)$

Previous works have shown that PageRank is biased towards old papers in scientific citation networks (Chen et al., 2007; Mariani et al., 2015; Maslov & Redner, 2008). Moreover, we have shown in Section 3 that PageRank score p is biased by scientific domain. To simultaneously suppress these two biases, we propose the age- and field-rescaled PageRank score $R^{AF}(p)$. $R^{AF}(p)$ is defined similarly as $R^{AF}(c)$: we compute the mean value $\mu_i^{AF}(p)$ and the standard deviation $\sigma_i^{AF}(p)$ of the PageRank scores of the papers that belong to the same field as paper i and that are among the Δ_p closest papers to i as measured by the distance $|i-j|$ between their rank by age. The age- and field-rescaled PageRank score is then defined as

$$R_i^{AF}(p) = \frac{p_i - \mu_i^{AF}(p)}{\sigma_i^{AF}(p)}. \tag{5}$$

In the following, we set $\Delta_p = 1000$.

4.3. Field bias of the new metrics

In the top panels of Fig. 3, we show that in the top-1% of the rankings by $R^{AF}(p)$ and $R^{AF}(c)$ each field appears well represented. In fact, the deviations from the expected value are very small especially if compared to the deviations of the other rankings (see top panels in Figs. 1 and 2). In the bottom panels of Fig. 3, we report that the full score distributions for papers from different fields collapse extremely well top of each other thanks to the rescaling procedure.

5. Quantifying rankings' biases by field and age

We begin this Section by introducing a new methodology to assess a ranking's bias based on the Mahalanobis distance (Section 5.1). Then, we use this to quantify the bias by field (Section 5.2) and the bias by age and field (Section 5.3).

5.1. A general framework to assess ranking biases based on the Mahalanobis distance

While Figs. 1 and 2 illustrate the substantial field bias of the existing metrics, the bias is much weaker (if any) for the new age- and field-rescaled metrics in Figure 3 . Now we quantify this improvement by extending the statistical tests of bias suppression presented by Mariani et al. (2016), Radicchi and Castellano (2012b), Radicchi et al. (2008), Waltman et al. (2012). Similarly to these works, we assume that a ranking is unbiased if its properties are consistent with those of an unbiased selection process.

5.1.1. Assessing the bias by field

Let us first analyze the problem of assessing the bias by field. Consider an urn which contains N marbles, each of them corresponding to one of the publications present in our dataset. An *unbiased selection process* then corresponds to sampling from this urn at random without replacement a fixed number $n = \lfloor N \times 0.01 \rfloor$ of publications. From the extracted sample, we count the number of publications that belong to each field f, k_f, and record these numbers in the vector $\vec{k} = (k_1, \ldots, k_F)^T$; here

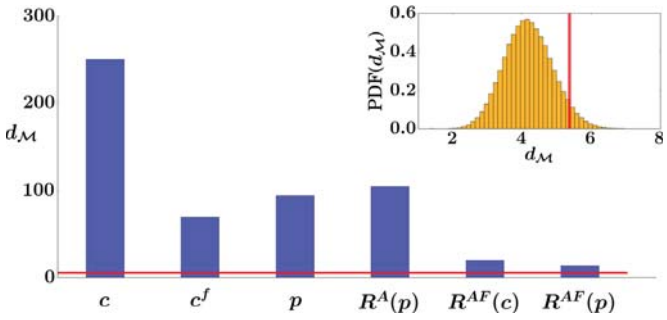

Fig. 4. Mahalanobis distances, $d_{\mathcal{M}}$, for the analyzed indicators when considering the 19 main fields. From left to right: citation count, relative citation count, PageRank, age-rescaled PageRank, age- and field-rescaled citation count and age- and field-rescaled PageRank. The horizontal red line represents the upper bound of the 95% confidence interval obtained from the simulations. In the insets, we report the distribution of $d_{\mathcal{M}}$ coming from 1 000 000 simulations of the unbiased sampling process. Again, the red line represents the upper bound of the 95% confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

F denotes the number of fields. The probability to observe a certain vector, $\vec{k}$, is given by the *multivariate hypergeomentric distribution* (MHD)

$$P\left(\vec{k}\right) = \frac{\prod_f^F \binom{K_f}{k_f}}{\binom{N}{n}} \tag{6}$$

where K_f is the total number of publications in field f. Following this selection process, among the n extracted publications, the expected number of publications for field f is $\mu_f = n\,K_f/N$.

Assume that the actual ranking by a given metric m features $k_f^{(m)}$ publications from field f in the top 1% of its ranking. In general, the observed $k_f^{(m)}$ deviates from its expected value μ_f. A simple approach to quantify this deviation would consist in computing the z-score, defined as $z_f^{(m)} := (k_f^{(m)} - \mu_f)/\sigma_f$, where σ_f is the expected standard deviation for field f according to the MHD specified by Eq. (6). There are however two shortcomings of the z-score. First, the z-score only gives partial information for a MHD – how far from the expected values we are in units of standard deviations – but it does not provide information on how statistically significant the deviations are. Second, to quantify the overall bias of a given indicator m, we would need to aggregate the z-scores from the different fields. For example, we could take the average z-score, but this would neglect the correlation between the different fields coming from the constraint $n = \sum_f k_f^{(m)}$.

To overcome these two problems, we follow a different approach. We first run various numerical simulations that reproduce the unbiased selection process. These simulations produce a set of ranking vectors which are distributed according to Eq. (6) around the vector of expected values, $\vec{\mu} = (\mu_1, \ldots, \mu_m)$. Differently from (Radicchi & Castellano, 2012b), we do not estimate the confidence interval for the different fields separately. We calculate instead the Mahalanobis distance ($d_{\mathcal{M}}$, Mahalanobis (1936) and Appendix B) for each simulated vector from $\vec{\mu}$, and construct the distribution of $d_{\mathcal{M}}$'s obtained by the simulated unbiased selection process. The inset of the left panel of Fig. 4 reports the distribution of the $d_{\mathcal{M}}$ for 1 000 000 simulations. The distribution is centered around its mean value of 4.18 and the upper bound for the 95% confidence level is around 5.37. [2] For an *ideal unbiased ranking*, we would expect its $d_{\mathcal{M}}$ to fall into the 95% confidence interval of the distribution of the $d_{\mathcal{M}}$ obtained from the simulated unbiased sampling process.

5.1.2. *Assessing the bias by age and field*

The methodology presented above is easily generalized to simultaneously assess a ranking's bias by age and field.

To add the temporal dimension to the bias assessment procedure, we split the publications into T equally-sized age groups, and repeat the above analysis by using $F \times T$ different categories of publications. In Section 5.3, we set $F = 19$ representing the number of fields and $T = 40$ as in (Mariani et al., 2016), and thus we obtain 760 age-field groups of different sizes.

[2] A curiosity for the reader. Here, the average of the square of the $d_{\mathcal{M}}$ for the unbiased sampling process is extremely close to the number of degrees of freedom of our problem. This stems from the fact that the MHD defined by Eq. (6) converges to a Multivariate Gaussian Distribution (MGD) as we increase the number of publications N while keeping n/N fixed and small. Our dataset is large enough for this approximation to be accurate. The $d_{\mathcal{M}}^2$ of a MGD is distributed as a χ^2 variable with average equal to the number of degrees of freedom, i.e. 18 since we have 19 distinct fields and one constraint.

Table 1
The individual contribution $z_i^2\,(1-k_i/N)$ of each field i to the $d_{\mathcal{M}}$ of the different metrics.

Field	c	c^f	p	$R^A(p)$	$R^{AF}(c)$	$R^{AF}(p)$
Art	1.15	1.95	3.01	0.31	2.84	0.09
Biology	**36.46**	15.81	6.74	0.87	0.12	0.16
Business	2.06	2.08	5.95	1.51	14.56	13.50
Chemistry	0.34	8.44	0.85	9.28	4.23	0.00
Computer Science	0.29	**23.86**	0.02	3.09	0.12	13.50
Economics	3.34	6.51	4.20	5.77	**32.32**	7.93
Engineering	8.36	0.09	10.48	0.35	8.36	0.03
Environmental Science	16.82	0.04	**35.67**	29.89	2.45	2.23
Geography	0.05	1.34	0.15	0.50	0.15	0.51
Geology	1.02	3.04	6.42	0.13	2.10	10.06
History	0.69	2.48	1.67	0.15	3.12	0.52
Materials Science	0.39	0.43	0.23	0.01	2.37	0.10
Mathematics	5.17	1.94	0.10	0.14	0.09	**25.34**
Medicine	4.02	7.66	0.06	1.94	2.16	19.56
Philosophy	1.49	3.23	0.12	0.36	0.15	0.07
Physics	8.30	0.21	6.00	**33.67**	14.34	0.01
Political Science	1.47	10.22	6.82	0.51	1.47	2.44
Psychology	5.75	1.43	8.56	10.24	2.93	1.65
Sociology	2.83	9.23	2.95	1.30	6.11	2.30

The bold values mark the fields that give the largest contribution to the $d_{\mathcal{M}}$ of each metric.

5.2. Results on the bias by field

The rankings produced by different metrics differ greatly by their $d_{\mathcal{M}}$ (see Fig. 4). As expected, the metrics that are not field-rescaled (c, $R^A(p)$, and p) are far from being unbiased. At the same time, relative citation count that is rescaled by field performs only slightly better than PageRank which is ignorant of any field information. The best results by a wide margin are achieved by our metrics, $R^{AF}(c)$ and $R^{AF}(p)$, obtained using the new rescaling. Nevertheless, both these metrics fail to meet the 95% upper bound achieved by simulated unbiased rankings. As disappointing as it may seem, this finding is not entirely surprising as the proposed rescaling procedures focus on equalizing the first two moments of the respective quantities (c and p) whereas the quantities' distributions can differ also by higher moments.

To understand which field contributes the most to the resulting $d_{\mathcal{M}}$ values, we derived an alternative analytic expression for the $d_{\mathcal{M}}$

$$d_{\mathcal{M}}(\vec{k},\vec{\mu})^2 = \sum_i^F z_i^2\left(1-\frac{k_i}{N}\right) \tag{7}$$

where we omit the metric superscript (m) from the notation for z_i and $\vec{k}$ for simplicity. We have proven this formula analytically for $F=3, 4, 5, 6$, and we have numerically tested it for $F=19$ and 760 (see Appendix B); it remains open to prove it in arbitrary dimensions.

In Table 1, we report the individual fields' contributions to the $d^2_{\mathcal{M}}$ calculated using Eq. (7). We find that Biology and Computer Science are the fields which give the biggest contributions to the $d^2_{\mathcal{M}}$ for the rankings by citation count and relative citation count, respectively. This could not have been detected by looking at the deviations from the expected values. Indeed, in Fig. 1 we only see that Environmental Science and Political Science have the largest deviations. For the novel indicators, approximately one third of the $d^2_{\mathcal{M}}$ of $R^{AF}(c)$ is explained by the field of Economics and approximately one fourth of the $d^2_{\mathcal{M}}$ of $R^{AF}(p)$ is explained by the field of Mathematics.

In addition, we also find that the $d_{\mathcal{M}}$'s contributions across different fields assume values in a relatively broad range. This suggests that findings on rankings' bias by field may strongly depend on which disciplines are included or not in the analysis. We argue that arbitrary choices on which fields to include should be avoided in future research on field-normalization of impact indicators.

To summarize, our bias suppression test allows us not only to estimate the level of bias ($d_{\mathcal{M}}$) of the various metrics, but also to quantify which percentage of the total bias ($d^2_{\mathcal{M}}$) of an indicator is explained by each single field.

5.3. Results on the bias by age and field

While the analysis of the previous section focused on the ranking bias by field, in this section we use the $d_{\mathcal{M}}$ to simultaneously assess the bias by age and field of a given ranking.

In Fig. 5, we show the $d_{\mathcal{M}}$'s for the different indicators and for the 95% confidence interval for the simulated unbiased selection process using 40×19 age-field types of publications. For citation count, PageRank, relative citation count and age-rescaled PageRank we have to reject the hypothesis that the rankings of these indicators are not biased by age and field. For

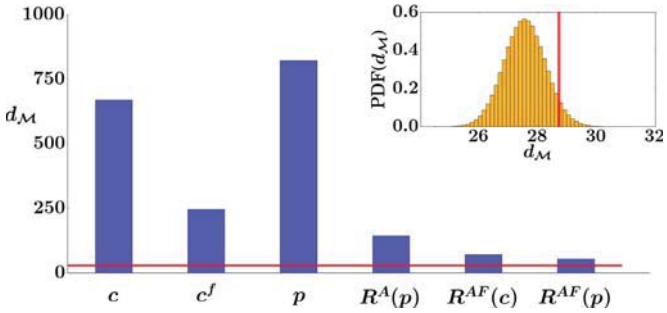

Fig. 5. Mahalanobis distances, $d_{\mathcal{M}}$, for the analyzed indicators when considering the 760 age-field groups. From left to right: citation count, relative citation count, PageRank, age-rescaled PageRank, age- and field-rescaled citation count and age- and field-rescaled PageRank. The horizontal red line represents the upper bound of the 95% confidence interval obtained from the simulations. In the insets, we report the distribution of $d_{\mathcal{M}}$ coming from 1 000 000 simulations of the unbiased sampling process. Again, the red line represents the upper bound of the 95% confidence interval. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

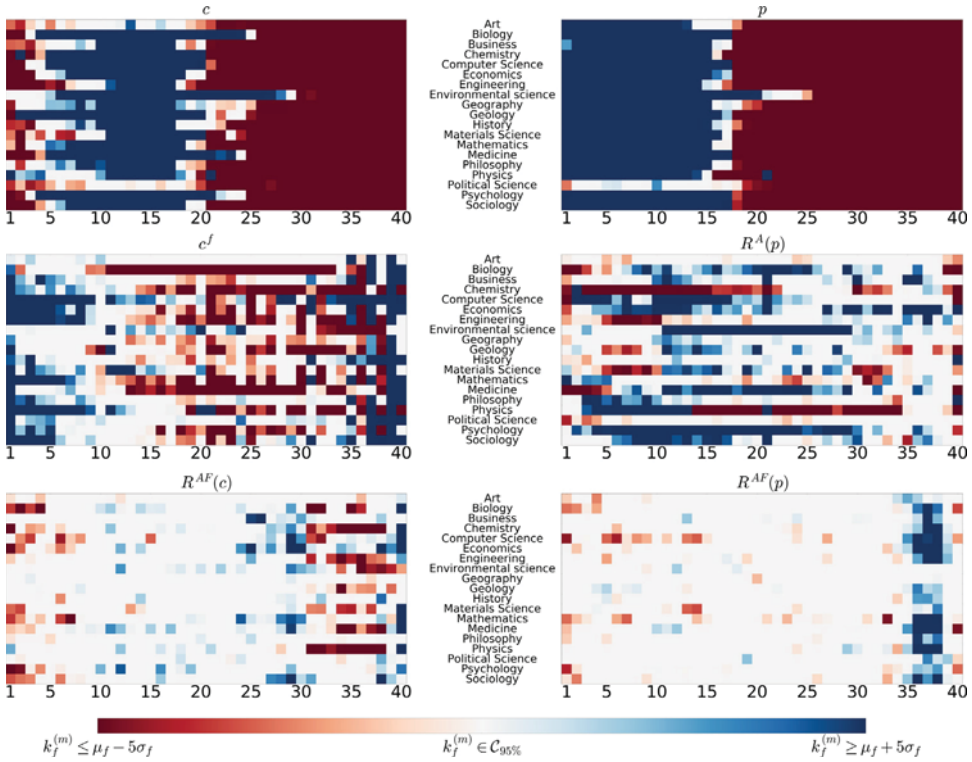

Fig. 6. Heatmaps showing the bias by field and age of the rankings by the different indicators. Each cell represents an age-field group: age groups are represented horizontally, while fields are represented vertically. The color of the cells shows the bias of the indicators with respect to that age-field group. White means that the respective age-field group is fairly represented in the top 1% of the ranking by the indicator. While we use a color scale from white to intense red (blue) for age-field group which are underestimated (overestimated). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

the improved indicators, age- and field-rescaled citation count and PageRank, we also have to reject the null hypothesis, even though they are much closer to the 95% confidence interval.

It is worth to notice that age-rescaled PageRank, an indicator developed to only remove PageRank's bias by age (Mariani et al., 2016), is significantly less biased compared to relative citation count, an indicator specifically designed to simultaneously remove bias by age and field.

5.4. Simultaneously visualizing the bias by age and field

To visualize the field and age bias of the rankings by the analyzed indicators, we use heatmaps in the age-field group plane (see Fig. 6). In these heatmaps, each cell represents a field-age group, and its color indicates the level of bias. A white cell indicates that the number of papers in the respective age-field group falls into the 95% confidence level ($\mathcal{C}_{95\%}$) determined

with the simulations. Hence, white means that no bias is detected for that age-field group. While for representing the bias towards or against a group of papers, we use blue (overestimation) and red (underestimation). To obtain a range of over/under-estimation, the brightness of the colors ranges from white (no bias) to intense blue/red. The most intense colors indicate that the number of papers from that age-field group is 5 standard deviation smaller/bigger than the expected value.

The top panel of Fig. 6 shows that, independently of field, citation count and PageRank systematically over-represent old papers and under-represent recent papers. This is in agreement with the findings of several other works (Chen et al., 2007; Mariani et al., 2015, 2016; Newman, 2009). The only exception is Political Science which is usually underestimated independently of paper age. We argue that this happens because this is the smallest field in the data set, and it has become an academic discipline by itself much later compared to most of the other fields[3]. Also, the oldest papers in most fields are under-represented by citation count, which reflects the change of citation practices over time.

The middle panels of Fig. 6 show that the relative citation count and age-rescaled PageRank suppress large part of the biases of the original metrics, yet specific fields are consistently overestimated or underestimated. For example, both age-rescaled PageRank and relative citation count under-represent papers belonging to the field of Chemistry. A peculiarity of relative citation count is that it over-represents both the oldest as well as the most recent papers at the cost of the other papers.

The bottom panels of Fig. 6 show the heatmaps for the new indicators age- and field- rescaled citation count $R^{AF}(c)$ and PageRank $R^{AF}(p)$. We find that the respective rankings are much less biased towards specific fields compared to all the other analyzed measures. However, there are two patterns: for $R^{AF}(c)$ recent publications tend to be underestimated for some fields, whereas for $R^{AF}(p)$ recent publication tend to be overestimated for almost all fields. These rather systematic patterns must have their roots in changes of the citation and PageRank score distributions with time. Since our rescaling procedure was fixing the first two moments of these distributions, the observed patterns come from differences in higher moments. Thus, the distributions of $R^{AF}(c)$ and $R^{AF}(p)$ are aligned only partially for papers of different age.

6. Discussion and conclusion

To summarize, in this paper we have analyzed a large citation network from the Microsoft Academic Graph to show that the rankings of papers by well-known indicators are extremely biased by age and field. The level of bias of the rankings has been quantified with a new statistical framework based on the Mahalanobis distance. This framework has allowed us to simultaneously quantify the age and field biases of the analyzed rankings, and to determine which groups of papers give the largest contributions to the observed bias. To allow other researcher to easily implement our statistical test for ranking bias, we make the respective code publicly available[4] together with a quick tutorial on how to use it.[5] In addition, we have also introduced two new indicators of paper impact, rescaled citation count $R^{AF}(c)$ and rescaled PageRank $R^{AF}(p)$ that produce much less biased rankings than existing indicators. In particular, the ranking by $R^{AF}(p)$ is approximately three times less biased compared to the least biased existing metrics, relative citation count and age-rescaled PageRank.

The contribution of our results to the debate on the validity of field-normalization procedures is threefold. First, our findings are in agreement with the conclusions of Albarrán et al. (2011) and Waltman et al. (2012) which argued that the relative citation count introduced by Radicchi et al. (2008) can be insufficient to effectively remove citation count's bias by age and field. Second, we show the importance of testing indicators using an accurate statistical procedure, such the one introduced in this paper. Indeed, for the least-biased indicators analyzed, $R^{AF}(c)$ and $R^{AF}(p)$, no clear indication of bias is found at first glance. However, when using the statistical test based on the Mahalanobis distance, we find a significant discrepancy between their rankings and those coming from unbiased sampling process. We argue that including higher-order momenta (such as the skewness) in the rescaling procedure can be an efficient way to further reduce the rankings' level of bias. Third, by deriving an explicit formula to calculate the contribution of each field to the bias of a ranking, we find that the these contributions assume a broad range of values. We obtain similar findings also for the contributions to the age-field bias. This means that the level of bias of rankings depends heavily on which years or fields are included in the analysis. For this reason, in future research on age and field normalization of indicators, it is essential to clearly motivate which years and fields are included in the analysis, avoiding arbitrary or uncritical decisions.

To address the bias by age and field of ranking of papers, we have first divided the papers in groups with similar age and from the same field. Then, we considered only the sizes of these groups to define an unbiased selection process from which we obtained a statistical null model for an unbiased ranking. In principle, additional information can be included into the null model to correct for other effects. For example, including information about the co-authorship network would permit to correct for the effect of this network on the growth and structure of the citation network (Sarigöl, Pfitzner, Scholtes, Garas,

[3] We notice that the classification of Political Science as one of the highest-level fields is not obvious. In Scopus categories, "Political Science and International Relations" is only a subfield of the higher-level field Social Science [http://www.scimagojr.com/journalrank.php?area=3300]. In the Web of Science classification scheme, "Political Science" is only a subfield of the higher-level field "Social Sciences, General" [http://ipscience-help.thomsonreuters.com/inCites2Live/8300-TRS.html].

[4] https://github.com/giava90/quantifying-ranking-bias.

[5] https://www.sg.ethz.ch/team/people/gvaccario/quantifying-ranking-bias/.

& Schweitzer, 2014; Sarigol, Garcia, Scholtes, & Schweitzer, 2017). In this way, we would gain a better understanding of how the social dimension of science contributes to the field and age biases of impact indicators.

We emphasize that removing the biases addressed in this paper and those that come from social aspects is of primary importance not only for scholarly publication databases, but also for several other information systems, such as the WWW or online social networks (Scholtes, Pfitzner, & Schweitzer, 2014a). As a matter of fact, every day scholars and on-line users explore available knowledge using recommender systems based on ranking algorithms. This challenges us to design more sophisticated filtering and ranking procedures to avoid biases that can systematically hide relevant contents or only show information too similar to what the users already know.

While we have analyzed in detail the presence of age and field bias in the ranking, it still remains to evaluate the actual ranking performance of the newly proposed indicators in artificial data (Medo & Cimini, 2016) or in real data where the ground truth is provided by some external source (Dunaiski, Visser, & Geldenhuys, 2016; Mariani et al., 2016). Another important issue is the comparison between metrics based on citation count and metrics that take the whole citation network into account to determine papers' score. Our analysis (see Section D) shows that the rankings by the least-biased indicators $R^{AF}(c)$ (citation-based) and $R^{AF}(p)$ (network-based) are positively correlated, still substantially different. Can we use the extra-information provided by the network to enhance our ability to identify highly-significant publications? Our intuition and the results presented by Chen et al. (2007) and by Mariani et al. (2016) for specific research fields suggest that this is the case. Yet, we need additional analysis to validate this conjecture in larger datasets such as the one analyzed in this article.

To conclude, by reducing the age and field biases from indicators of scientific impact and by extending the existing statistical tests for biases, we contribute to the challenge of quantifying and suppressing biases of rankings in information systems.

Acknowledgements

We thank Ingo Scholtes, Frank Schweitzer and Yi-Cheng Zhang for suggestions which improved the manuscript. In addition, we also thank Emre Sarigöl for his help in pre-processing the data, and Elias Bauman for his important contribution to the optimization of the C++ code that we used for the network analysis. GV acknowledges support from the Swiss State Secretariat for Education, Research and Innovation (SERI), Grant No. C14.0036 as well as from EU COST Action TD1210 KNOWeSCAPE. MSM acknowledges support from the Swiss National Science Foundation Grant No. 200020-156188.

Author contributions Conceived and designed the analysis: Giacomo Vaccario, Matus Medo, Nicolas Wider, Manuel Sebastian Mariani.

Collected the data: Giacomo Vaccario.

Contributed data or analysis tools: Giacomo Vaccario, Manuel Sebastian Mariani.

Performed the analysis: Giacomo Vaccario, Manuel Sebastian Mariani.

Wrote the paper: Giacomo Vaccario, Matus Medo, Nicolas Wider, Manuel Sebastian Mariani.

Appendix A. KDD Cup data

A.1 Data source

In this work, we analyzed the dump of the *Microsoft Academic Graph* (MAG) released for the KDD Cup 2016 (Sinha et al., 2015), a competition linked to a prestigious computer science conference on knowledge discovery and data mining (KDD). Among the primary interests of the community organizing the KDD Cup, there are the technical challenges related to web-scale data collection and aggregation. For this reason, the released data for the KDD Cup 2016 went through only basic processing.[6] Each publication in the dataset is endowed with its unique identifier, *paperid*, *publication date*, *references* to other publications, and its *field of study*.

A.2 Data pre-processing

When analyzing the data, we do not distinguish the publications by their type (paper, review, book, etc.). Further, we also do not differentiate between different types of journals and take into account all of them: for example, we do not distinguish between a citation coming (or going) from (or to) a letter or a book. We argue that it is important to keep various types of journals and publications because different fields adopt not only different citation norms, but also different ways to communicate their results. For example, computer science researchers commonly publish results in conference proceedings, while physics authors tend to prefer articles or letters. At the same time, we are aware that different types of publications might have different citation characteristics. However, good indicators should ideally be able to account for heterogeneities among publications and citation norms across different communities and produce unbiased rankings without the need for

[6] https://kddcup2016.azurewebsites.net/Data

Table A.2
Main fields. The 19 parent main fields identified by Microsoft Academic Graph with their number of publications and average citations. The second to last row reports the total number of publications considering multiple times publications that belong to more than one fields, whereas the last row reports the total number of unique publications and the average citation count.

Field	Publication count	Mean citation count
Art	233 251	3.90
Biology	5 847 554	9.67
Business	613 827	4.81
Chemistry	6 204 531	7.28
Computer Science	4 080 636	6.13
Economics	2 252 921	5.56
Engineering	3 011 763	5.10
Environmental Science	315 465	12.63
Geography	288 338	6.92
Geology	1 825 707	7.88
History	390 144	5.53
Materials Science	2 063 474	6.12
Mathematics	4 551 453	5.87
Medicine	5 061 990	7.90
Philosophy	787 649	5.05
Physics	6 976 644	5.55
Political Science	144 473	2.51
Psychology	2 861 813	8.23
Sociology	1 784 695	5.39
Total	49 296 327	–
Total (no multiple)	18 193 082	6.42

arbitrary choices about which types of articles to include in the analysis. In addition, similarly as Waltman et al. (2012) and differently from Radicchi et al. (2008), we do not exclude publications which do not receive citations.

As mentioned in Section 2, the MAG has a field classification scheme with 4 hierarchical levels. The field assignment is based on an internal algorithm that uses a machine learning approach (Sinha et al., 2015). In our work, similarly to Radicchi et al. (2008), we are only interested in impact metric normalization at the most coarse-grained level. To this aim, in our analysis, we focus only on the 19 main fields as listed in Table A.2. Discussing the possible limitations of the classification approach by MAS and the dependence of our results on the adopted classification scheme is a relevant subject for future research but goes beyond the scope of this manuscript. We only included in the analysis papers for which the following information are available: (1) unique identifier (ID); (2) complete publication date (yyyy/mm/dd) crucial for the temporal rescaling procedure that is explained in the following; (3) DOI or journal-id, in order to be able to retrieve the publication; (4) assignment to at least one of the main 19 fields. We discard from our analysis publications for which one or more of these four properties are missing.

With this filtering procedure, we obtain $N = 18\,193\,082$ unique publications and $E = 109\,719\,182$ citations.

A.3 Data basic statistics

We first observe that the distributions of both incoming and outgoing edges are broad (see Fig. A.7). Both these distribution in fact have long and heavy tails. In addition, we also observe strong variations of the mean citation count across fields (see Table A.2 for details). For example, publications belonging to the field of Environmental Science and of Political Science have

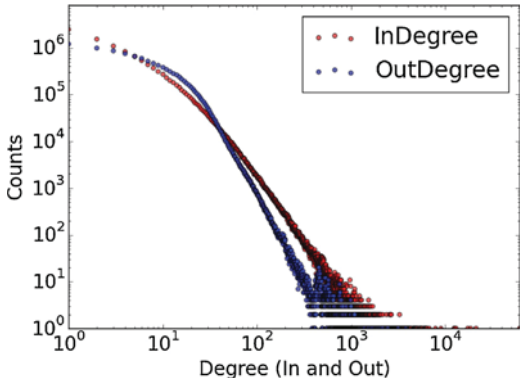

Fig. A.7. In- and out-degree distribution for the publications present in our dataset after preprocessing the data.

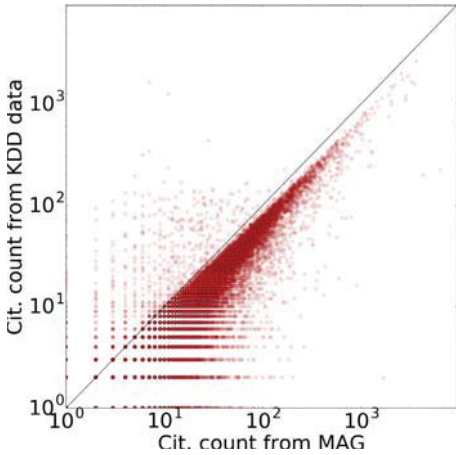

Fig. A.8. Scatter plot of the citation counts reported in the data released for the KDD Cup 2016 and from the online version of the MAG (02/2017).

mean citation count two times bigger and smaller, respectively, compared to the average citation count across all fields. In agreement with the findings by Radicchi et al. (2008), Schubert and Braun (1986), Vinkler (1986), the strong variety in mean citation count per publication for the various fields confirms that the corresponding communities exhibit different citation behavior, which calls for field normalization procedures.

A.4 Data quality

Here, we address the following question: to which extent is the KDD dump of the MAG in accordance with the most updated online version of the MAG?[7] For this comparison, we divided the 49 296 327 analyzed paper-field pairs (one paper-field pair is composed of a paper and the fields that paper belongs to) into 760 age and field groups, as described in Section 5. From each group, we randomly choose 0.1% of papers. Following this procedure, we obtained a representative sample with respect to field and age composed of approximately 50 000 papers.

First, we have matched the paper ID in hexadecimal format present in the data released for the KDD Cup to the paper ID in int64 format present in the on-line version of the MAG. For 50 papers, we manually verified that the paper IDs were exactly the same in the two datasets, albeit represented in different formats. Then, using the Academic Knowledge Api,[8] we have downloaded the number of citations for each sampled paper.

For the 2.3% of the sampled papers, we were not able to find their corresponding paper in the online MAG. For 50 unmatched papers, we found out that these were papers with duplicates in the KDD Cup data, i.e. these papers are present in the KDD data with two distinct IDs, and only one of the two IDs is present in the MAG.

For the matched papers (97.7% of the sample), we compared the number of citations reported in the KDD dump of the MAG with the number of citations reported on the online version of the MAG. Since the online MAG has a longer time span than the KDD data, in the absence of noise, we would expect the number of citations in the KDD data to be smaller than or equal to the number of citations reported in the online MAG. We find that the two citation counts are highly correlated (Fig. A.8), and only the 2.2% of the sampled papers have more citations in the KDD data compared to the online version of the MAG. This sets a lower boundary for the error in the percentage of papers with wrong number of citations to about 2.2%.

To summarize, after our filtering procedure, we find that the data released for KDD Cup 2016 has about 2.3% of papers with duplicates. In addition, about 2.2% of the matched papers have errors in their citation count. This means that we have correct citation information for about 95.6% of the analyzed papers.

Appendix B. Evaluating individual contributions to the Mahalanobis distance

The Mahalanobis distance ($d_{\mathcal{M}}$) is an established measure in statistics which generalizes the concept of z-score to multivariate distributions by taking into account also possible correlations between the random variables (Mahalanobis, 1936). Its definition reads

$$d_{\mathcal{M}}(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \mathbf{S}^{-1} (\vec{x}-\vec{y})} \tag{B.1}$$

[7] We have performed this analysis during February 2017.
[8] https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api

where $\mathbf{S}^{-1}$ is the inverse of the covariance matrix, $\vec{x}$ and $\vec{y}$ are two vectors containing the random variables. When the covariance matrix is diagonal, i.e. the random variables are not correlated, than the $d_{\mathcal{M}}$ is equivalent to the square root of the sum of the squares of the z-scores.

In Section 5.2, we have used Eq. (7), an expression for the $d_{\mathcal{M}}$ valid when the covariance matrix comes from a Multivariate Hypergeometric Distribution (MHD), i.e., when the elements of the matrix are

$$S_{ij} = \left(\delta_{ij}(K_i(N-K_i)) - \left(1-\delta_{ij}\right)K_iK_j\right)\gamma \;\forall i,j = 1,\ldots,F-1 \tag{B.2}$$

where δ_{ij} is the Kronecker delta, K_i is the number of papers of category i, $N = \sum_i^F K_i$ is the total number of papers, F is the number of paper categories, $\gamma = n(N-n)/N^2(N-1)$ and n is the number of sampled papers. It is worthy to remember that even though we have F different categories, we only have $F-1$ degrees of freedom. Here, we derive Eq. (7) for the case of a MHD in three dimensions, i.e., for $F=3$. In this case, the covariance matrix is 2×2:

$$\mathbf{S} = \gamma\begin{pmatrix} K_1(N-K_1) & -K_1K_2 \\ -K_1K_2 & K_2(N-K_2)\end{pmatrix} \tag{B.3}$$

and the inverse of the covariance matrix is

$$\mathbf{S}^{-1} = \frac{1}{\gamma\det(\mathbf{S})}\begin{pmatrix} K_2(N-K_2) & K_1K_2 \\ K_1K_2 & K_1(N-K_1)\end{pmatrix} \tag{B.4}$$

where $det(\mathbf{S}) = K_1(N-K_1)K_2(N-K_2) - (K_1K_2)^2$ denotes the determinant of the covariance matrix, S. Then, let us consider two random column vectors extracted from a 3-dimensional MHD, $\vec{x} = (x_1,x_2,x_3)^T$ and $\vec{y} = (y_1,y_2,y_3)^T$ such that $n = \sum_{i=1}^3 x_i = \sum_{i=1}^3 y_i$ where n is the number of sampled papers. Substituting Eq. (B.4) in Eq. (B.1), we write the square of the $d_{\mathcal{M}}$ between $\vec{x}$ and $\vec{y}$ as

$$\begin{aligned} d_{\mathcal{M}}(\vec{x},\vec{y})^2 = & \frac{1}{\gamma\det(\mathbf{S})}\begin{pmatrix} x_1-y_1 & x_2-y_2\end{pmatrix}\begin{pmatrix} K_2(N-K_2) & +K_1K_2 \\ +K_1K_2 & K_1(N-K_1)\end{pmatrix}\begin{pmatrix} x_1-y_1 \\ x_2-y_2\end{pmatrix} \\ = & \frac{1}{\gamma\det(\mathbf{S})}\Big\{(x_1-y_1)^2K_2(N-K_2) + (x_2-y_2)^2K_1(N-K_1) \\ & + 2(x_1-y_1)(x_2-y_2)(K_1K_2))\Big\} \\ = & \frac{1}{\gamma\det(\mathbf{S})}\Big\{(x_1-y_1)^2K_2(K_1+K_3) + (x_2-y_2)^2K_1(K_2+K_3) \\ & + 2(x_1-y_1)(x_2-y_2)(K_1K_2)\Big\} \\ = & \frac{1}{\gamma\det(\mathbf{S})}\Big\{(x_1-y_1)^2K_2K_3 + (x_2-y_2)^2K_1K_3 \\ & + [(x_1-y_1)+(x_2-y_2)]^2K_1K_2\Big\} \end{aligned}$$

where we have used $N = \sum_{i=1}^3 K_i$. Recalling that $n = \sum_{i=1}^3 x_i = \sum_{i=1}^3 y_i$, we know that $(x_1-y_1)+(x_2-y_2) = (x_3-y_3)$, so we write

$$d_{\mathcal{M}}(\vec{x},\vec{y})^2 = \frac{1}{\gamma\det(\mathbf{S})}\left\{(x_1-y_1)^2K_2K_3 + (x_2-y_2)^2K_1K_3 + (x_3-y_3)^2K_1K_2\right\} \tag{B.5}$$

Then, by using the relation $\det(\mathbf{S}) = N\prod_{i=1}^3 K_i$, we have:

$$d_{\mathcal{M}}(\vec{x},\vec{y})^2 = \frac{1}{\gamma}\sum_{i=1}^3\frac{(x_i-y_i)^2}{NK_i}; \tag{B.6}$$

noticing from Eq. (B.2) that $\gamma = S_{i,i}/(K_i(N-K_i))$, we obtain

$$d_{\mathcal{M}}(\vec{x},\vec{y})^2 = \sum_{i=1}^3\frac{(x_i-y_i)^2}{S_{ii}}\frac{K_i(N-K_i)}{NK_i} = \sum_{i=1}^3\frac{(x_i-y_i)^2}{S_{ii}}\left(1-\frac{K_i}{N}\right) \tag{B.7}$$

Finally, if we choose one of the two vectors to contain the expected values, μ_i, we re-obtain Eq. (7) since $(x_i-\mu_i)^2/S_{ii} = z_i^2$. To be precise, the covariance matrix is not defined for $i=3$, however the relation $\gamma = \sigma_3^2/(K_3(N-K_3))$ holds and therefore also the final result.

Using Mathematica or similar softwares, it is easy to prove analytically that eq. (7) holds for small dimensions. We have verified it until 6 dimensions. Moreover, we have numerically tested this formula by calculating the $d_{\mathcal{M}}$'s between the ranking vectors of the indicators and the vector of expected values, $\vec{\mu}$, with two different alternative methods: (1) by using Eq. (B.1),

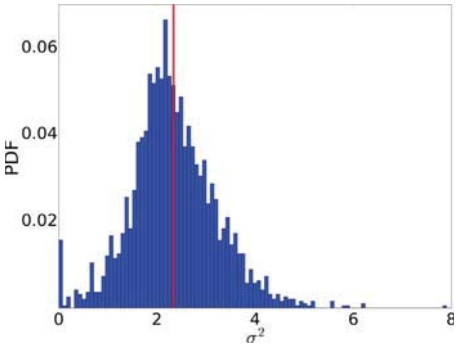

Fig. C.9. Distribution of σ^2 obtained by calculating the empirical ratio, $r = (e^{\sigma^2} - 1)$, between the variance and the square of the mean citation count of each field and year.

i.e., by inverting the covariance matrix, and (2) by using the eigenvalue decomposition of the covariance matrix.[9] The results of the three methods were all compatible with each other up to 10 decimal digits. The advantage of using Eq. (7) is that we can calculate the $d_{\mathcal{M}}$ between two arbitrary vectors without dealing with any (computationally slow) matrix inversion or diagonalization, and the number of needed calculations scales linearly with the number of dimensions. Importantly, Eq. (7) allows also to assess the individual contribution of each dimension (i.e. of each category) to the $d^2_{\mathcal{M}}$. To our best knowledge, we are the first ones to have derived such explicit formula for $d_{\mathcal{M}}$ when the covariance matrix and the random vectors come from a MHD.

Appendix C. First moment rescaling

According to Radicchi et al. (2008), the distribution of the c^f indicator is log-normal:

$$F(c^f)dc^f = \frac{1}{\sigma c^f \sqrt{2\pi}} e^{-[\log(c^f)-\mu]^2/2\sigma^2} dc^f \tag{C.1}$$

where $\mu = -\sigma^2/2$ and σ is fitted from the data. When Eq. (C.1) is verified, then also the distributions of citation count, c_i, for all the individual fields, i, are lognormal:

$$F(c_i)dc_i = \frac{1}{\sigma c_i \sqrt{2\pi}} e^{-[\log(c_i)-\log(c_0)-\mu]^2/2\sigma^2} dc_i \tag{C.2}$$

where c_0 is the mean of c_i. For lognormal distributions the variance is proportional to the square of the mean and the constant of proportionality is $(e^{\sigma^2} - 1)$. From Eq. (C.2), we see that the citation counts c_i are distributed lognormally with mean $e^{\mu+\log(c_0)+\sigma^2/2}$ and variance $(e^{\sigma^2} - 1)e^{2\mu+2\log(c_0)+\sigma^2}$. Recalling that $\mu = -\sigma^2/2$, we have that the mean is c_0, as it is expected, while the variance becomes $(e^{\sigma^2} - 1)c_0^2$. Thus, when eq. (C.1) is verified, the variance of the empirical distribution of the citations for each field has to be proportional to the square of the mean citation count. Moreover, the constant of proportionality has to be $(e^{\sigma^2} - 1)$ for every field and year.

The analytic result just presented is in line with Eq. (C.1) given in Appendix C of Mariani et al. (2016). There it is shown that a rescaling procedure based on diving the original score by their first moment works if the ratio between standard deviation and mean is constant. In the case of the relative citation ratio, we can calculate analytically such constant using the lognormal distribution and obtain the fitting parameter σ^2.

In Fig. C.9, we report the distribution of σ^2 obtained by calculating the empirical ratio between the variance and the square of the mean, r, and by inverting the relation $r = (e^{\sigma^2} - 1)$ for every field and year. If the universality claim was correct, we would expect a narrow distribution of σ^2. By contrast, we find that σ^2 ranges between 0 and 8 across different fields and years. We argue that the broad range of σ^2 is the reason why the first moment rescaling introduced in (Radicchi et al., 2008) does not work in the analyzed dataset.

[9] The matrix $\mathbf{S}$ is symmetric and it has maximal rank because it is the covariance matrix of a multivariate distribution. Therefore, we can diagonalize it, $\mathbf{S} = \mathbf{B}^{-1}\mathbf{D}\mathbf{B}$ where the columns of $\mathbf{B}$ form an orthonormal basis; we can also write $\mathbf{S}^{-1} = \mathbf{B}^{-1}\mathbf{D}^{-1}\mathbf{B}$. With this, we have $d^2_{\mathcal{M}}(\vec{x}, \vec{y}) = \sum_i^{F-1} c_i/\lambda_i$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{S}$ and $\{c_i\}$ are the coordinates of $\vec{x} - \vec{y}$ in the basis which diagonalizes $\mathbf{S}$, i.e. $c_i = \sum_k^{F-1}(x_k - y_k)B^{-1}_{ki} = \sum_k^{F-1}(x_k - y_k)B_{ik}$ where the last equality comes from the orthonormality of $\mathbf{B}$ which implies $\mathbf{B}^{-1} = \mathbf{B}^T$.

Table D.3
Correlations between the metrics. The four columns represent (from left to right): Pearson's correlation coefficient r, r restricted to papers that received at least ten citations, Spearman's rank correlation coefficient ρ, and ρ restricted to papers that received at least ten citations.

Metrics	r	r (only $c > 10$)	ρ	ρ (only $c > 10$)
c, p	0.82	0.82	0.79	0.59
$R^{AF}(c)$, $R^{AF}(p)$	0.80	0.81	0.82	0.74

Appendix D. Comparing the rankings by the metrics

In Section 5, we have shown that the rankings by citation count and PageRank can be substantially leveled off across different fields and age groups through a rescaling procedure based on the z-score. The two resulting indicators, $R^{AF}(p)$ and $R^{AF}(c)$, are the least biased among the indicators considered in this paper. It is important to notice that while citation count has a direct interpretation and it is widely used in research evaluation (Waltman, 2016), PageRank score is a more sophisticated quantity which has not yet be turned into a standard tool for research assessment. If the rankings by rescaled PageRank and rescaled citation count brought similar information, rescaled citation count might be preferred due to its simpler interpretation and easier computation.

Differently from the citation count, PageRank score uses information on the whole network topology to compute each paper's score. While there exists no universal criterion to decide whether PageRank score leads to a better ranking than the citation count, the results by Chen et al. (2007) and Mariani et al. (2016) suggest that in citation networks, PageRank score improves our ability to find groundbreaking publications in the data. On the other hand, a node that received many citations is more likely to achieve larger PageRank score – Fortunato, Bogu ná, Flammini, and Menczer (2006) showed that PageRank score is on average proportional to citation count for uncorrelated networks.

We address now the question: *To what extent the rankings by (rescaled) PageRank and (rescaled) citation count differ?* We focus on two correlations: (1) the correlation between citation count and PageRank, which has been of interest for previous studies (Chen, Gan, and Suel (2004), Fortunato et al. (2006), Pandurangan, Raghavan, and Upfal (2002)) due to the essential role played by PageRank algorithm in determining the success of Google's Web search engine; (2) the correlation between the two rescaled indicators $R^{AF}(c)$ and $R^{AF}(p)$. The measured correlations are all positive, significantly larger than zero, yet significantly smaller than one (see Table D.3). This is interesting as it may point out that, in analogy with the findings by Chen et al. (2007), Mariani et al. (2016), Ren, Mariani, Zhang, and Medo (2017), network topology brings useful information that is neglected by citation count. Whether the additional information used by PageRank metric can be used to identify groups of significant papers will be the subject of future research.

Differently than the rankings by citation count and PageRank score (not shown here), the top papers identified by $R^{AF}(c)$ and $R^{AF}(p)$ come from diverse historical period and diverse fields. Due to the lack of time bias, also very recent papers can reach the top of the rankings by $R^{AF}(c)$ and $R^{AF}(p)$. It would be instructive to look at the top papers as ranked by rescaled citation count and rescaled PageRank. However, the original MAG datasets presents noisy entries, as reported by the KDD cup 2016 (see Appendix A for more details), and it causes the scores of some recent papers to be over-estimated, which makes some of the entries in the top-20 of the rankings by the rescaled scores unreliable. For this reason, we do not show the rankings here. At the same, this problem does not affect the statistical results presented in the previous Sections.

References

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zooma test of zitts hypothesis. *Scientometrics*, *75*(1), 81–95.
Albarrán, P., Crespo, J. A., Ortu no, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics*, *88*(2), 385–397.
Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The eigenfactor metrics. *Journal of Neuroscience*, *28*(45), 11433–11434.
Bonacich, P. (1987). Power and centrality: A family of measures. *American Journal of Sociology*, *92*(5), 1170–1182.
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., & Daniel, H.-D. (2009). Universality of citation distributions – A validation of radicchi et al.'s relative indicator $c_f = c/c_0$ at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology*, *60*(8), 1664–1670.
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, Y.-Y., Gan, Q., & Suel, T. (2004). Local methods for estimating pagerank values. *Proceedings of the thirteenth ACM international conference on information and knowledge management, ACM*, 381–389.
Chen, P., Xie, H., Maslov, S., & Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics*, *1*(1), 8–15.
Colliander, C., & Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics*, *5*(1), 101–113.
de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306.
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.
Ermann, L., Frahm, K. M., & Shepelyansky, D. L. (2015). Google matrix analysis of directed networks. *Reviews of Modern Physics*, *87*(4), 1261.
Fortunato, S., Bogu ná, M., Flammini, A., & Menczer, F. (2006). Approximating PageRank from in-degree. In *International workshop on 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. (2006). The history and meaning of the journal impact factor. *JAMA*, *295*(1), 90–93.
Gleich, D. F. (2015). Pagerank beyond the web. *SIAM Review*, *57*(3), 321–363.
Harzing, A.-W., & Alakangas, S. (2013). Microsoft academic: Is the phoenix getting wings? *Scientometrics*, 1–13.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, *1656*, 9–16572.
Hug, S. E., & Braendle, M. P. (2017). *The coverage of Microsoft academic: Analyzing the publication output of a university.* , arXiv preprint arXiv:1703.05539
Hug, S., Ochsner, M., & Brandle, M. P. (2017). Citation analysis with Microsoft academic. *Scientometrics*, *111*(1), 371–378.
Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, *18*(1), 39–43.
Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, *46*(5), 604–632.
Lundberg, J. (2007). Lifting the crowncitation *z*-score. *Journal of Informetrics*, *1*(2), 145–154.
Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, *4*, 9–55.
Mariani, M. S., Medo, M., & Zhang, Y.-C. (2015). Ranking nodes in growing networks: When PageRank fails. *Scientific Reports*, *5*.
Mariani, M. S., Medo, M., & Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics*, *10*(4), 1207–1223.
Maslov, S., & Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *Journal of Neuroscience*, *28*(44), 11103–11105.
McAllister, P. R., Narin, F., & Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management*, *4*, 205–211.
Medo, M., & Cimini, G. (2016). Model-based evaluation of scientific impact indicators. *Physical Review E*, *94*(3), 032312.
Medo, M., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters*, *107*, 238701.
Newman, M. E. J. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)*, *86*(6), 68001.
Pandurangan, G., Raghavan, P., & Upfal, E. (2002). Using PageRank to characterize web structure. In *International computing and combinatorics conference*. pp. 330–339. Springer.
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). Why Sirtes's claims (Sirtes, 2012) do not square with reality. *Journal of Informetrics*, *6*(4), 615–618.
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.
Ren, Z.-M., Mariani, M. S., Zhang, Y.-C., & Medo, M. (2017). *A time-respecting null model to explore the structure of growing networks.* , arXiv preprint arXiv:1703.07656.
Sarigöl, E., Pfitzner, R., Scholtes, I., Garas, A., & Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science*, *3*(1), 1.
Sarigol, E., Garcia, D., Scholtes, I., & Schweitzer, F. (2017). Quantifying the effect of editor-author relations on manuscript handling times. *Scientometrics*, 1–25 (in press).
Scholtes, I., Pfitzner, R., & Schweitzer, F. (2014). The social dimension of information ranking: A discussion of research challenges and approaches. In *Socioinformatics – The social impact of interactions between humans and IT*. pp. 45–61. Springer.
Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C. J., & Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications*, *5*.
Schubert, A., & Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, *9*(5-6), 281–291.
Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. P., et al. (2015). An overview of Microsoft academic service (MAS) and applications. *Proceedings of the 24th international conference on world wide web, ACM*, 243–246.
Sirtes, D. (2012). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics*, *6*(3), 448–450.
van Leeuwen, T. N., & Medina, C. C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation*, *21*(1), 61–70.
Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics*, *10*(3-4), 157–177.
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., & van Eck, N. J. (2010). The relation between eigenfactor, audience factor, and influence weight. *Journal of the American Society for Information Science and Technology*, *61*(7), 1476–1486.
Waltman, L., Yan, E., & van Eck, N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics*, *89*(1), 301–314.
Waltman, L., van Eck, N. J., & van Raan, A. F. J. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology*, *63*(1), 72–77.
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391.
Yao, L., Wei, T., Zeng, A., Fan, Y., & Di, Z. (2014). Ranking scientific publications: The effect of nonlinearity. *Scientific Reports*, *4*, 6663.
Zhang, Z., Cheng, Y., & Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of web of science subject categories. *Scientometrics*, *101*(3), 1679–1693.
Zhou, J., Zeng, A., Fan, Y., & Di, Z. (2016). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics*, *106*(2), 805–816.
Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401.