TimeRank: A dynamic approach to rate scholars using citations

Massimo Franceschet
Department Mathematics, Computer Science, and Physics
University of Udine – Italy
massimo.franceschet@uniud.it

Giovanni Colavizza
Digital Humanities Laboratory
École Polytechnique Fédérale de Lausanne – Switzerland
giovanni.colavizza@epfl.ch

September 11, 2017

Abstract

Rating has become a common practice of modern science. No rating system can be considered as final, but instead several approaches can be taken, which magnify different aspects of the fabric of science. We introduce an approach for rating scholars which uses citations in a dynamic fashion, allocating ratings by considering the relative position of two authors *at the time of the citation* among them. Our main goal is to introduce the notion of *citation timing* as a complement to the usual suspects of popularity and prestige. We aim to produce a rating able to account for a variety of interesting phenomena, such as positioning raising stars on a more even footing with established researchers. We apply our method on the bibliometrics community using data from the Web of Science from 2000 to 2016, showing how the dynamic method is more effective than alternatives in this respect.

1 Introduction

For better or worse, modern scientists have become accustomed with the quantification of their performance and its use for recognition, such as funding allocation [Wildgaard et al., 2014]. A variety of indicators exist, relying on different systems to rate scholars and, quite consequently, rank them. Several indicators rely on citations. In general, every rating system is an attempt at highlighting a specific aspect which might be of interest when considering scholars. The two most important aspects factored into citation-based indicators are popularity, or the number of endorsements, and prestige, or the rank of endorsers [Bollen et al., 2006; Franceschet, 2010].

An aspect of importance has nevertheless been overlooked. Since science is a cumulative effort where every contribution is published at a certain time, the resulting citation network is a dynamic and open-ended process. More specifically, citations happen at a given time, when a certain relation in terms of relative rating exists among authors, which could be different at another point in time. Yet the most established rating systems for scholars are *static* methods that disregard the dynamic nature of the underlying process. In so doing, we argue, they filter out meaningful information on *the timing of citations*.

The timing of citations allows to give a premium to scholars who are cited by higher rated scholars against the odds, at a time when they would be unlikely to be cited by them. Consider, for example, the case of two researchers who both have received endorsements from the most accomplished scholars in their field, thus have similar quality of citations, yet one is at an early career stage, the other is an already senior scholar who received such endorsement only late in his career. The former scholar has a timing advantage, having received early citations against the odds, the latter has a quantitative advantage, having had a longer career, accumulating more publications and citations. Using static, time-insensitive, indicators it is likely that the latter scholar would score higher, yet by considering the timing of their recognition, the resulting rating would even out. As another example, consider two researchers receiving a citation from the same scholar but in different periods. The first researcher is endorsed when the citing scholar is yet an unknown author, while the second researcher is endorsed when the citing scholar has become a popular and esteemed author. Again, static methods do not acknowledge the difference in the timing of the two citations, since citations come from the very same author. On the other hand, a dynamic approach accounts for this difference.

The method we propose updates ratings with citation rewards computed sequentially in time, by considering the relative ratings of the scholars involved in the endorsement when a citation is given. At the end of the process, the method outputs a time series of citations rewards for each scholar: the rating of a scholar is defined as the sum of all his citation rewards. As a result, popularity (number of endorsements), prestige (rank of endorser and endorsed) and timing (time of endorsement) are all accounted for in the final rating. The analysis of such citation time series can further help to distinguish between similarly rated scholars with different citation reward histories, e.g. rising versus consolidated or declining researchers. Indeed recent literature on detecting rising stars found temporal features to be the most discriminative for this task [Zhang et al., 2016]. We think our method can be best applied in a situation when the dynamics of a scholar's performance should be considered alongside its quantitative and qualitative aspects. Examples are the early identification of rising stars to help hiring committees in their decisions, or the distinction among stable or declining scholars at any level of seniority. The method in fact enriches the very notion of reputation in a field by putting on a more even footing scholars of different career stages, that have contributed substantially and have been recognised by their highest-ranking peers.

We start by outlining the dynamic rating method, its main properties and variants. We then discuss its application to the bibliometrics community, by considering a dataset from the Web of Science for the years 2000 to 2016. The article closes with a discussion of related works and some final remarks.

2 TimeRank: a dynamic rating method

Consider a temporal citation network among scholars: nodes are scholars and an edge (i, j, t) is a citation from scholar i to scholar j at time t. This means that there is a paper authored by i and published at time t that cites a paper authored by j. TimeRank, the time-sensitive method we propose in this paper, works as follows:

1. initially, at time 0, all scholars have the same rating, say 0;

2. then, citations are processed in increasing temporal order. At any time $t > 0$, the ratings of scholars cited at time t are simultaneously updated in terms of their previous ratings at time $t - 1$ and the previous ratings at time $t - 1$ of the citing scholars. If a scholar is not cited at time t then his rating does not change.

More specifically, let $i_1, \ldots, i_n$ be the (possibly not unique) scholars citing j at time $t > 0$. The rating r_j of scholar j is updated using the following update rule:

$$r_j \leftarrow r_j + \sum_{k=1}^{n} \rho_{i,j} \tag{1}$$

where

$$\rho_{i,j} = \frac{10^{(r_i - r_j)/\zeta}}{1 + 10^{(r_i - r_j)/\zeta}} \tag{2}$$

is the citation reward that scholar j gains because he has been cited by scholar i and $\zeta > 0$ is a constant. The reward is expressed using the logistic curve depicted in Figure 1. Notice that:

- the reward $0 < \rho_{i,j} < 1$. In particular, it is always positive, hence all citations, even from bottom-ranked scholars, give a contribution to the rating of the cited scholar;

- the reward is high, close to 1, if the rating of the citing scholar i is significantly higher than that of the cited scholar j;

- the reward is low, close to 0, if the rating of the citing scholar i is significantly lower than that of the cited scholar j;

- the reward is intermediate, close to 0.5, if citing and cited scholars have similar ratings.

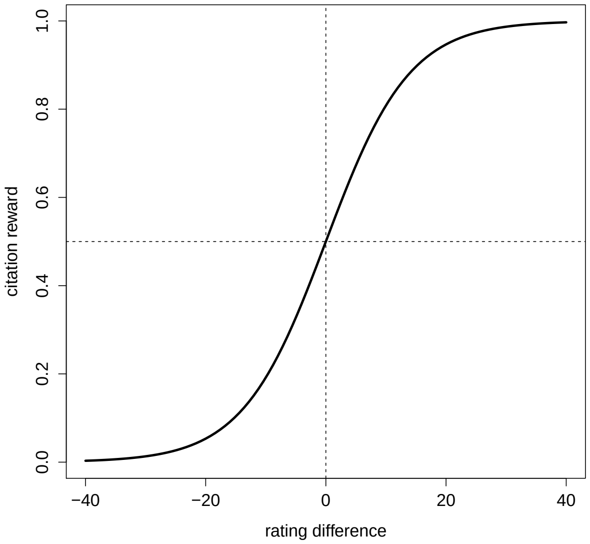

Figure 1: The citation reward in terms of the rating difference between citing and cited scholars ($\zeta = 16$).

2.1 Interpretation

The citation reward of Equation 2 can be rephrased as follows:

$$\rho_{i,j} = \frac{10^{r_i/\zeta}}{10^{r_i/\zeta} + 10^{r_j/\zeta}}.$$

It follows that:

$$\rho_{i,j} + \rho_{j,i} = 1$$

For instance, suppose $r_i - r_j = \zeta > 0$, so that there are ζ rating points of advantage of scholar i over scholar j. Then $\rho_{i,j} = 10/11 > \rho_{j,i} = 1/11$. Hence, the reward for j by receiving a citation from the higher rated scholar i is much larger (ten times larger) than the reward for i by receiving a citation from the lower rated scholar j. This leads to an interpretation of the role of the parameter ζ. We have that

$$\frac{\rho_{i,j}}{\rho_{j,i}} = \frac{10^{r_i/\zeta}}{10^{r_j/\zeta}}$$

and thus

$$\rho_{i,j} = \rho_{j,i} 10^{(r_i - r_j)/\zeta}$$

This means that for every ζ rating points of advantage that scholar i has over scholar j, the reward for scholar j, when cited by scholar i, is expected to be 10

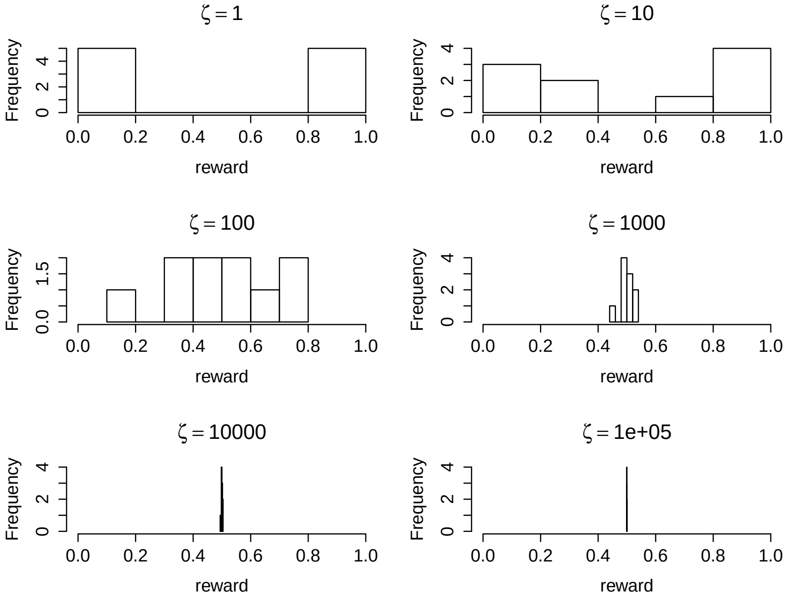

Figure 2: The distribution of rewards at increasing values of the parameter ζ. Rewards have been generated using the logistic function of Equation 2 on a uniform random sample of rating differences in the interval $(-100, 100)$.

times the reward scholar i would get, when cited by scholar j. It follows that, if two scholars have significantly different ratings, they have at least ζ rating points of difference.

The reward for a citation from an equal rated scholar is always 0.5. As ζ grows to ∞, it is more and more difficult to obtain rewards far from the intermediate value of 0.5 and, in the limit, all rewards are equal to 0.5. In this case the TimeRank ratings correspond to half the number of received citations, and hence there is perfect correlation with the number of citations. On the other hand, as ζ goes to 0, it is increasingly easier to gain rewards far from the intermediate value of 0.5. In the limit, all rewards assume three values: 0 for citations from lower rated scholars, 0.5 for citations from equal rated scholars, and 1 for citations from higher rated scholars. In this case the TimeRank ratings correspond to the number of citations received from higher rated scholars plus half of the number of citations received from equal rated scholars. See Figure 2 for a simulation of this effect varying the parameter ζ. Both these extremes are not interesting, because they simply count the number of rewards, without weighting them. An intermediate value for the parameter ζ – not too large, not too small – is hence reasonable.

2.2 Comparison with static methods

To better understand the dynamics of TimeRank, it is useful to analytically compare it with the total number of received citations (TotCit for short) and PageRank, which are the closest siblings in the context of bibliometric indicators. First, notice that the TimeRank rating of a scholar can be expressed as the sum of the citation rewards received by the scholar. Indeed, using Equation 1, we have that:

$$r_j = \sum_i \rho_{i,j} \tag{3}$$

where the sum is defined on all citations from some scholar i to scholar j and the reward $\rho_{i,j}$ is defined with respect to the ratings of scholars i and j *at the time of citation*. It follows that there is a correlation between TimeRank and TotCit: all else equal, if scholar i receives more citations than scholar j, then the rating of i is higher than the rating of j with both methods. Nevertheless, while TotCit simply counts citations, TimeRank *weights* them with a function of the spread of the ratings of the citing and cited scholars. In fact, TotCit is a special case of TimeRank, given by doubling the rewards in the limit when ζ goes to infinity (when rewards are all equal to 0.5). As for the PageRank method, recall that it considers three factors: (1) the number of citations received; (2) the ratings of the citing scholars; and (3) the citation propensity of citing scholars. TimeRank differs with respect to PageRank in two ways:

1. PageRank considers the absolute rating of the citing scholar, while TimeRank considers the *relative rating* of the citing scholar with respect to the cited scholar;

2. PageRank uses the ratings of citing scholars at the end of the temporal citation process (the same time for all citations), while TimeRank incorporates the *timing* of citations by using the ratings of citing and cited scholars at the actual time of citation (different times for different citations).

Hence, *relative rating and timing* are two original ingredients of TimeRank. We illustrate the importance of these factors in the example depicted in Figure 3 and in Table 1. We compared TimeRank with TotCit and PageRank. We used $\zeta = 16$ in TimeRank and a damping factor of 0.85 in PageRank. Consider the ratings of scholars C and D. Both scholars C and D receive a unique citation from A and B, respectively, but with different timing. In particular, C is cited by A when A is important (he received many citations), while D is cited by B when B is unknown (he received no citations). Both TotCit and PageRank methods have a static approach − they do not consider the temporal evolution of citations − and cannot distinguish between the positions of C and D. On the other hand, TimeRank has a temporal perspective and more reasonably favours C over D.

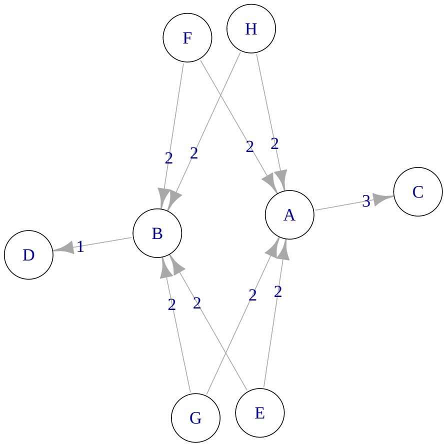

Figure 3: A simple temporal scholar citation network. The edges are stamped with the time of the corresponding citation.

2.3 Consistency

We discuss in this section the consistency of the proposed method. In general, by consistency [Waltman and Van Eck, 2009] or independence [Marchant, 2009a,b] it is meant a stability in the relative inequality of ratings between two scholars receiving identical change in conditions. For example, a consistent rating between scholars x and y, where $r_x < r_y$, would maintain the inequality of ratings after an identical change in both authors' conditions with respect to the rating method. In TimeRank, the setting is as follows. Suppose three authors are given, x, y and z, and $r_x < r_y$ at time $t-1$. Then, at time t, author z cites once both x and y. The method is consistent if, after updates, the inequality $r_x < r_y$ still holds.

We denote with $\hat{r}_x$ and $\hat{r}_y$ the ratings of x and y after the citation from z. Hence:

$$\begin{aligned}\hat{r}_x &= r_x + \tfrac{10^{r_z/\zeta}}{10^{r_z/\zeta}+10^{r_x/\zeta}} \\ \hat{r}_y &= r_y + \tfrac{10^{r_z/\zeta}}{10^{r_z/\zeta}+10^{r_y/\zeta}}\end{aligned}$$

We define a consistency function:

$$c(r_x, r_y, r_z, \zeta) = \hat{r}_y - \hat{r}_x = r_y - r_x + \frac{10^{r_z/\zeta}}{10^{r_z/\zeta} + 10^{r_y/\zeta}} - \frac{10^{r_z/\zeta}}{10^{r_z/\zeta} + 10^{r_x/\zeta}}$$

Scholar	**TimeRank**	**TotCit**	**PageRank**
A	2.00	4	16.89
B	2.00	4	16.89
C	0.57	1	20.61
D	0.50	1	20.61
E	0.00	0	6.25
F	0.00	0	6.25
G	0.00	0	6.25
H	0.00	0	6.25

Table 1: Comparing TimeRank, TotCit and PageRank for the citation network depicted in Figure 3.

Assuming $r_x, r_y, r_z \geq 0$, $r_x < r_y$, and $\zeta > 0$, consistency holds when $c(r_x, r_y, r_z, \zeta) > 0$. Figure 4 shows that there are some extreme cases in which consistency of the TimeRank method is violated. This happens when the ratings r_x and r_y of scholars x and y are very close and the parameter ζ is very small (close to 0). The rating r_z can be either smaller than r_x, or in between r_x and r_y, or larger than r_y. Recall that, as shown in Section 2.1, when ζ is small, rewards are close to 0 (if the citing author is lower in the ranking) or close to 1 (if the citing author is higher in the ranking).

In order to investigate the sign of the consistency function $c(r_x, r_y, r_z, \zeta)$ in a more analytical way we fix some of its parameters. Without losing generality, we assume $r_x = 0$, and set $r_y = \alpha > 0$. Moreover, we set $r_z = \alpha/2$, so that r_z is equidistant from r_x and r_y. As shown in Figure 4, this is the point of maximum violation of consistency (the minimum of the curve). Hence, the consistency function $c(r_x, r_y, r_z, \zeta)$ in four variables boils down to a simplified version $d(\alpha, \zeta)$ in only two variables:

$$d(\alpha, \zeta) = \alpha + \frac{1 - 10^{\alpha/2\zeta}}{1 + 10^{\alpha/2\zeta}}$$

Notice that $-1 < (1 - 10^{\alpha/2\zeta})/(1 + 10^{\alpha/2\zeta}) < 1$ and hence consistency is achieved, that is $d(\alpha, \zeta) > 0$, for all $\alpha \geq 1$. As for $\alpha < 1$, with some elementary algebra it holds that $d(\alpha, \zeta) > 0$ if and only if

$$\zeta > f(\alpha) = \frac{\alpha}{2 \log \frac{1+\alpha}{1-\alpha}}$$

For instance, if $\alpha = 0.5$, then $f(0.5) = 1/\log 81 \sim 0.52$. Hence, for all $\zeta > 1/\log 81$, the consistency holds. Notice that $f(\alpha)$ is a decreasing function between 0 and 1 and:

$$\lim_{\alpha \to 0} f(\alpha) = \tfrac{\ln 10}{4} \sim 0.58$$

$$\lim_{\alpha \to 1} f(\alpha) = 0$$

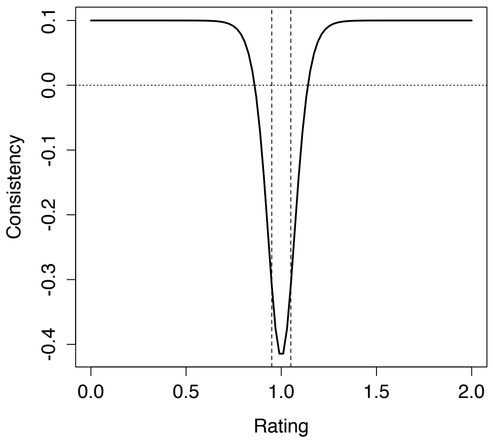

Figure 4: The violation of consistency. The ratings $r_x = 0.95$ and $r_y = 1.05$ (vertical lines), while r_z varies between 0 and 2 (x axis) and $\zeta = 0.1$. The part of the curve below zero corresponds to the ratings r_z that violate consistency.

Hence, for $\zeta > (\ln 10)/4$, consistency always holds (independently of α). In summary, the only inconsistent cases happen when the ratings r_x and r_y are very close and ζ is very small. As soon as ζ is sufficiently large, for instance $\zeta > 1$, the TimeRank ratings are consistent.

2.4 Variants of TimeRank

TimeRank is quite flexible. In this section we briefly describe some variations of the main theme. The interested reader can profitably refer to [Waltman, 2016] for a broader discussion of variants of citation impact indicators.

- *Self-citations.* In a self-citation, the citing and cited authors correspond, and hence the citation reward is always 0.5. A similar reward is earned if a scholar receives a citation from a similarly rated one. Assuming a long tail distribution of the ratings[1], the event of receiving a citation from a similarly rated scholar is more likely for low rated scholars than for high rated scholars. It follows that self-citations in this rating system implicitly reward high rated scholars more than low rated scholars. Furthermore, self-citations do not fit well with the approach taken, which rests on the difference in ratings of two scholars at the time of a citation between them. These might be good reasons to exclude self-citations in this rating system.

[1] The hypotheses is confirmed by our experiments.

- *Co-authorship.* Every citation from a paper with k authors to a paper with h authors, generates k citation rewards for every cited scholar. We can normalize this effect by dividing each citation reward by the number of authors of the cited paper, the citing paper, or both depending on the application.

- *Length of reference lists.* The method considers a citation to be one action, establishing a link between all combinations of citing and cited authors of two papers, for every reference in the reference list of the citing paper. It might be argued that the action should be the publication of the citing paper, as to filter out the fact that papers can have reference lists of variable length. We can normalize this effect by dividing the citation reward by the number of references made by the citing paper, following the fractional counting method [Leydesdorff et al., 2013; Perianes-Rodriguez et al., 2016].

- *Initial conditions.* Initially, all scholars are assigned an equal rating. This might be not realistic since, in general, authors have different potential or relative position at the beginning. One might use some exogenous factor to determine an initial rating of scholars and provide TimeRank with a hot instead of a cold start. The exogenous factor can be of any kind, qualitative or quantitative, provided it is internally consistent.

3 Case study

We propose an application of TimeRank on a subset of the bibliometrics community. We stress that the goal of this study is to better understand the details and nuances of TimeRank we propose, as well as to compare it with more traditional bibliometrics indicators.

We consider all articles published from (Jan) 2000 to (March) 2016 in the following journals: Scientometrics, the Journal of the Association for Information Science and Technology (including its previous relevant versions) and the Journal of Informetrics. We further consider only article typologies 'article' and 'review', for a total of 5952 individual publications (2831 Scientometrics, 579 Journal of Informetrics, 2542 Journal of the Association for Information Science and Technology). Citations indexed by the Web of Science are considered, as matched by the CWTS matching algorithms [Olensky et al., 2016]. We only keep citations to other articles within the dataset, in order to consider citations between authors that have published in these journals, thus that could be considered as part of the bibliometrics community.

Since we want to consider author to author citations, we use the CWTS author disambiguation method [Caron and van Eck, 2014], finding 7259 individual authors and 173509 citations among them (138507 excluding self-citations), by adding a citation among two authors if one cited an article authored by the other. This procedure naturally creates connections among all combinations of authors in multi-author publications. In our experiments we excluded self-citations, but did not apply any other normalisation procedure discussed in Section 2.4. Our dataset is timestamped by month. The rare forward citations were discarded, and ratings are updated synchronously (therefore synchronous citations are kept).

We implemented the TimeRank method in R [R Development Core Team, 2008], setting parameter $\zeta = 16$. The choice for ζ follows from the original proposal of the Elo's method, which inspired TimeRank (see Section 4). In Elo's method, by default parameters are $\kappa = 25$ and $\zeta = 400$ [Langville and Meyer, 2012], with κ commonly assuming values in between 10 and 32 in chess rating systems. Setting $\kappa = 1$ and $\zeta = 16$ removes one parameter (κ), without modifying the default ratio ζ/κ. We used the available implementations of PageRank from the igraph R package [Csardi and Nepusz, 2006], setting the damping factor to 0.85 (as in the original proposal of the method [Brin and Page, 1998]).

We performed the following three analyses:

1. an exploratory analysis, in which we describe the dataset with some basic statistics as well as compare TimeRank with traditional static bibliometric indicators;

2. a cluster analysis, with the aim of assigning scholars to performance classes based on their ratings;

3. a sensitivity analysis, with the goal of exploring the sensitivity of the TimeRank ratings to the timing of citations and to the variation of the parameter ζ.

3.1 Exploratory analysis

In this part we provide some descriptive statistics with the aim of exploring our dataset, including a comparison with traditional bibliometric indicators. The TimeRank ratings show the typical long-tail distribution with many low rated and few high rated scholars: 20% of the top-rated scholars accrue 70% of the total cumulative rating. The mean rating is 5.7, well above the median of 2.4; the maximum rating is 107.4. Two histograms of the distribution of the ratings are shown in Figure 5.

We compared the TimeRank method with TotCit, PageRank and the Hirsch index. There exists a positive correlation between TimeRank and these other bibliometric measures, as evident from Figure 6. However, this correlation is quite weak for highly rated scholars, see Table 2 and Figure 7 for a comparison of the ratings of top rated scholars.

To better illustrate the difference between TimeRank, TotCit and PageRank let us focus on the case of two scholars: LW and WG. Scholar WG leads both TotCit and PageRank rankings: he received 3240 citations and, according to the stochastic interpretation of PageRank, a random scholar would spend 2.139%

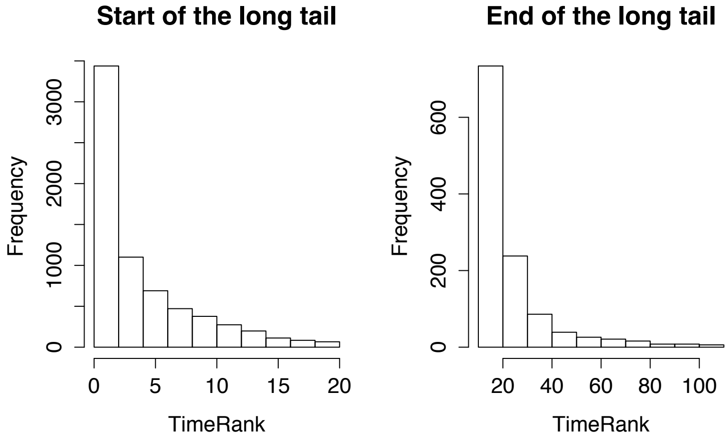

Figure 5: Histograms of TimeRank ratings. The left plot runs from the minimum value 0 to value 20, the right plot runs from values 10 to the maximum value 107.4.

Scholar	TimeRank	TotCit	PageRank	Hirsch
LW	107.45	1242	0.459	15
NJE	106.32	1141	0.438	14
AFJR	103.52	1757	1.335	16
WG	103.36	3240	2.139	21
TNL	102.61	1288	0.800	14
LL	100.72	2687	1.411	24
MSV	99.91	750	0.377	9
MZ	97.76	637	0.563	9
KWB	97.41	967	0.571	11
LB	94.91	1396	0.757	16

Table 2: The table shows the top-10 scholars ranked with respect to TimeRank, as well as the corresponding ratings for traditional bibliometric measures: TotCit, PageRank, Hirsch index.

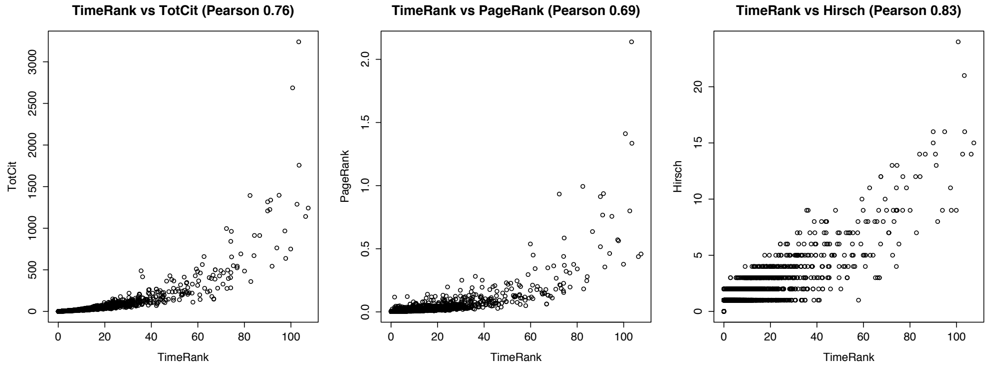

Figure 6: Scatterplots of TimeRank versus TotCit, PageRank and Hirsch index

of his time reading papers authored by him (out of 7126 other authors). Nevertheless, he is in 4th position in TimeRank ranking, with a rating of 103.36. The leader of TimeRank ranking is instead LW, with a rating of 107.45. He is, however, ranked 9th in the TotCit compilation, with 1242 citations (38% of the citations accrued by WG) and only 18th in the PageRank listing, with a score of 0.459 (22% of WG's score).

How can we explain these figures? Recall that the TimeRank of a scholar can be decomposed in the sum of rewards for citations he received from other scholars. With this in mind, notice that the mean reward of LW is 0.09 (with a standard deviation of 0.22), while the mean reward of WG is 0.03 (standard deviation is 0.09). The frequency of large rewards (greater than 0.5) is 7.2% for LW, corresponding to a share of 65.3% of his final rating. The frequency of large rewards is 0.5% for WG, corresponding to a share of 9.8% of his final rating. These numbers are given in Figure 8. All in all, despite the fact that WG received more citations, his rewards are smaller. On the other hand, LW obtained fewer but larger citation rewards. This leads to a TimeRank for the two scholars that is similar, with a little advantage for LW.[2]

3.2 Clustering scholars

While the TimeRank ratings provide a total order over scholars, it might happen that two or more scholars lie on different ranks but have similar ratings. The goal is then to group scholars in performance classes (or clusters), where a performance class is informally defined as a set of scholars with low intra-class rating distances and high inter-class rating distances.

[2] One might argue that these results are influenced by the fact that the dataset covers the period from 2000 to 2016, with WG publishing from 2001 and LW only from 2007. In fact, this is not the case. We repeated all experiments from 2007 and the outcomes do not significantly change.

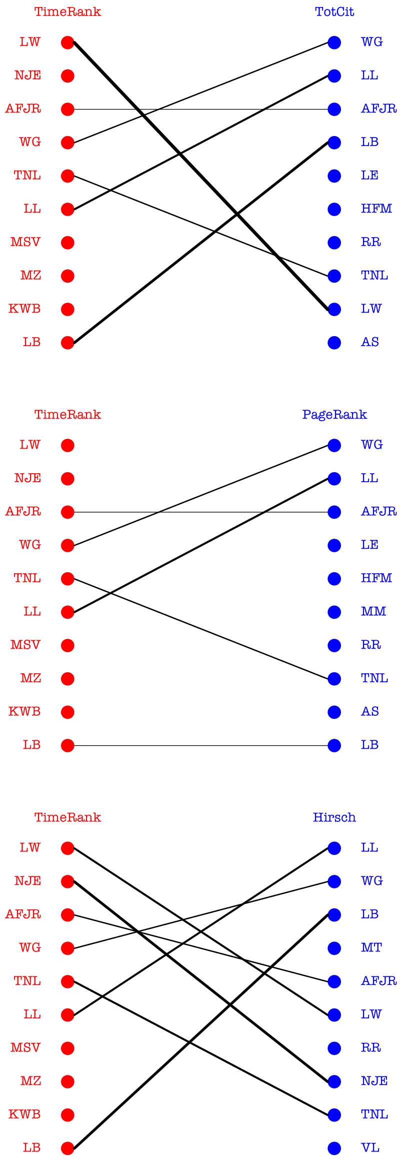

Figure 7: A visualization of top-10 rankings using TimeRank, TotCit, PageRank and Hirsch index. Lines connect the same scholar in the two rankings and line width is proportional to the rank displacement of the scholar.

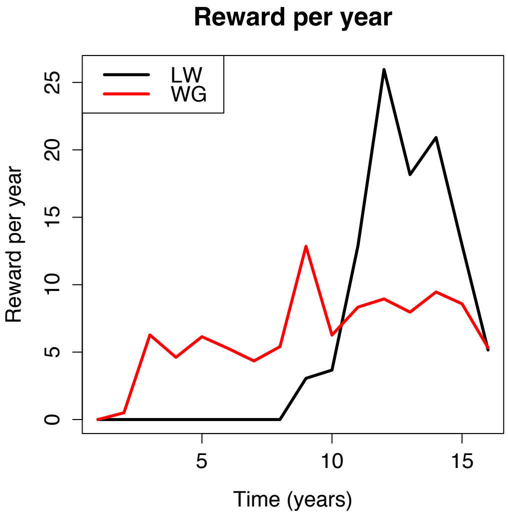

Figure 8: Reward per year for scholars LW and WG.

We used average-linkage hierarchical clustering to assign scholars to performance classes. The natural choice for the distance between two scholars is the absolute difference of the ratings of the scholars (the Euclidean distance). Since the clustering space has one dimension (all ratings can be arranged on a line), the resulting clusters of scholars are in fact contiguous intervals.

We first ran the clustering method on all the 7259 scholars, asking for k clusters, with k ranging from 1 to 12. The number 12 is the optimal number of clusters according to 4 indices of the NbClust R package [Charrad et al., 2014], assessed on a range from 2 to 15. Table 3 describes the resulting cluster intervals for all scholars. Notice that a couple of clusters are particularly robust: the top ranked scholars (those ranked up to position 21), and the bottom ranked scholars (in particular those ranked from position 774). Clusters tend to be smaller and more abundant in the top part of the ranking, meaning that this is the part of the ranking to which most of the rating variability can be traced. For instance, when we divide the ranking in 12 clusters, we have that 7 of them cover the top 120 scholars, and only 5 clusters cover the remaining 7139 authors.

In order to focus only on the top part of the ranking, we selected the top-84 rated scholars and repeated the clustering on them. Table 4 describes the resulting cluster intervals for top scholars, and Figure 9 depicts the cluster dendrogram. The optimal number of clusters suggested by the NbClust package is 3. The top cluster [1, 21] is still present, but it soon splits into smaller intervals, and when 10 clusters are found, it is decomposed into the following 5 clusters: [1,2], [3,5], [6,9], [10,16], and [17,21].

An alternative method to identify performance classes based on a ranking

1	–	–	–	–	–	–	–	–	–	–	–	7259
1	–	–	22	–	–	–	–	–	–	–	–	7259
1	–	–	22	–	–	–	–	183	–	–	–	7259
1	–	–	22	–	–	85	–	183	–	–	–	7259
1	–	–	22	–	–	85	–	183	–	774	–	7259
1	–	–	22	–	53	85	–	183	–	774	–	7259
1	10	–	22	–	53	85	–	183	–	774	–	7259
1	10	–	22	–	53	85	121	183	–	774	–	7259
1	10	–	22	–	53	85	121	183	449	774	–	7259
1	10	17	22	–	53	85	121	183	449	774	–	7259
1	10	17	22	37	53	85	121	183	449	774	–	7259
1	10	17	22	37	53	85	121	183	449	774	2404	7259

Table 3: Clusters (intervals) of all scholars. The table reads as follows. Scholars are sorted in decreasing order of rating and numbered from 1 to 7259, where 1 is the highest rated scholar and 7259 is the lowest rated scholar. Each line k, from 1 to 12, corresponds to a clustering of scholars into k contiguous intervals. Each value in a line is the index of the scholar that starts a new group. For instance, line 4 identifies the following 4 intervals of scholars: [1, 21], [22, 84], [85, 182], [183, 7259].

1	–	–	–	–	–	–	–	–	–	84
1	–	–	–	–	22	–	–	–	–	84
1	–	–	–	–	22	–	–	53	–	84
1	–	–	10	–	22	–	–	53	–	84
1	–	–	10	17	22	–	–	53	–	84
1	–	–	10	17	22	–	37	53	–	84
1	–	–	10	17	22	–	37	53	64	84
1	–	6	10	17	22	–	37	53	64	84
1	–	6	10	17	22	26	37	53	64	84
1	3	6	10	17	22	26	37	53	64	84

Table 4: Clusters (intervals) of top-84 scholars. See caption of Table 3 for details.

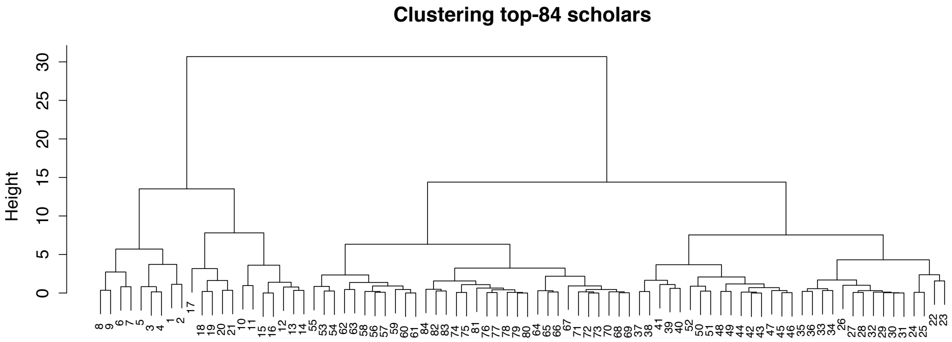

Figure 9: Cluster dendrogram for top-84 scholars.

is to use percentiles. The percentile-based method groups scholars by counting them. For instance, if we want to cluster the top-100 scholars into 10 clusters, the percentile-based method first sorts scholars in decreasing order of rating and then assigns the first 10 scholars to the 1st cluster, the second 10 scholars to the 2nd cluster, and so on. The resulting classes therefore all have the same number of members and, if scholar ratings are not homogeneously distributed on the rating line, they might contain scholars with significantly different ratings. On the other hand, the cluster-based method we have used takes into consideration the actual distance between scholar ratings. Hence, the resulting clusters might reflect substantively different performance classes.

3.3 Sensitivity

The most important original ingredient of TimeRank is its sensitivity to timing: the citing reward for j when he is cited by i is defined with respect to the ratings of both i and j *at the citation time*. On the other hand, static bibliometric indicators are not sensitive to the timing of citations. In this section we explore and quantify the sensitivity of TimeRank to timing.

To this end, we devised the following experiment. The dataset we use in our experiments can be represented as a table with three columns: citing scholar, cited scholar, and timestamp. By permuting the order of citations (first two columns), while the timestamp column is maintained fixed, we simulate a dataset with the same citations but with different timing of citations. Since all citations are exactly the same, static bibliometric indicators do not change their output on the simulated dataset. On the other hand, we expect TimeRank to be sensitive to timing and hence to compute a different output on a simulated dataset. We generated 1000 simulated datasets and computed

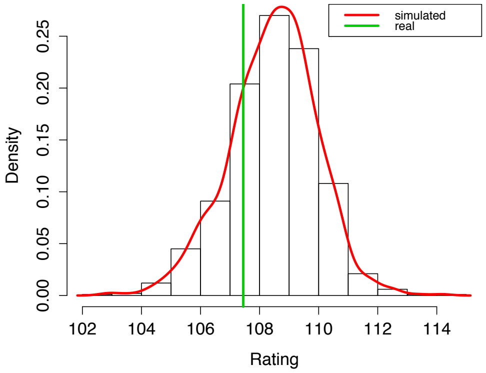

Figure 10: Histogram of simulated ratings for the top rated scholar. The red curve is the kernel density estimation of the simulated ratings while the green vertical line corresponds to the real rating.

the TimeRank ratings on each dataset. Hence, for each of the 7259 scholars, we have 1000 simulated ratings computed on the simulated datasets plus one rating computed on the real dataset.

Figure 10 depicts the histogram of the simulated ratings for one scholar (the leader of the ranking) as well as the real rating of the scholar. The distribution of the ratings follows a bell-shaped curve and is not centered around the real rating. If TimeRank were not sensitive to timing, then all simulated ratings would be equal to the real rating. Figure 11 shows the minimum, median and maximum of the simulated ratings for all scholars ranked from 1 to 100 (the top 100 scholars). Once again, if TimeRank were not sensitive to timing, then the three lines would coincide. On the contrary, the width of the band determined by the maximum and minimum values is on average 17% of the median, with a peak of 24% for the median. On the set of all scholars, the mean coefficient of variation of the simulated ratings is 4% with a peak of 17%. The mean distance between a typical simulated rating and the real rating is 10% of the real rating. We can conclude that, in general, TimeRank is significantly sensitive to timing, as intended.

Finally, we explore the sensitivity of TimeRank ratings to the parameter ζ. We have seen in Section 2 that extreme values of this parameter, that is values very close or very far from 0, are not meaningful. In our experiments, we used the intermediate value of 16. How sensitive are the ratings to values of ζ close to the chosen one? We computed the ratings for the following values of ζ: 2, 4, 8, 16, 32, 64, 128, and 256. It turns out that the sensitivity to changing the parameter is very low: the ratings with different values of the parameter are highly correlated; see Figure 12. Hence, any intermediate value of the parameter

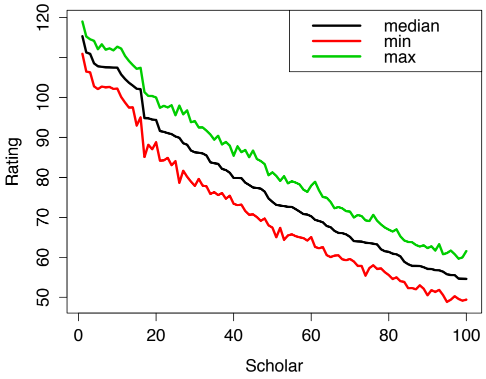

Figure 11: Minimum, median and maximum simulated ratings for the top-100 scholars.

ζ is suitable.

4 Related work

Author-level indicators have found widespread adoption in modern science [Wildgaard et al., 2014]. A set of simple output indicators are among the most used: the number of publications, the number of received citations or the mean number of received citations. Several indicators qualify outputs globally, for example using variants of the PageRank [Page et al., 1999]. Yet other indicators have been extensively adopted amidst some controversy, such as the Hirsch index [Hirsch, 2005; Waltman and van Eck, 2012]. In recent years, the great variety of proposed indicators has led the bibliometrics community to focus on their critical appreciation and practical use [Hicks et al., 2015].

PageRank and its variants are perhaps the indicators best suited to accommodate time dynamics [Fiala et al., 2008; Waltman and Yan, 2014; Radicchi et al., 2009; Yan and Ding, 2011; West et al., 2013]). The basic intuition of PageRank and its variants, that is to account for both the quantity of endorsements and the prestige of endorsers, has also been implemented by other methods (e.g. the AP-Rank [Zhou et al., 2012] or the P-Rank [Yan et al., 2011]).

Many real-world networks are in fact time-resolved: the date of each interaction between pairs of vertices, which forms an edge of the network, is recorded. Temporal networks, also called time-varying or dynamic networks, are graphs in which edges are labelled with temporal information about the relationship between the edge nodes. The static network analysis toolkit, including centrality measures, has been extended to dynamic networks [Holme and Saramäki, 2012; Nicosia et al., 2013; Holme, 2015]. Our work naturally embeds into this

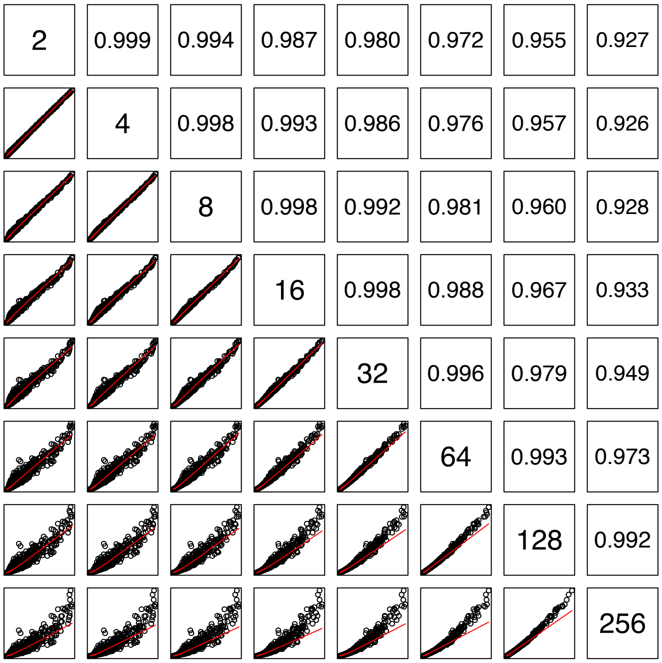

Figure 12: Scatterplot matrix of TimeRank ratings varying the parameter ζ. The matrix reads as follows. Each cell (i, j), with $i > j$ (lower part of the matrix) contains a scatterplot with regression line and the mirror entry (j, i) (upper part of the matrix) contains the corresponding Pearson correlation coefficient. The diagonal entries (i, i) and (j, j) contain the ζ parameters used for the computation.

context, since the scholar citation network we use is an instance of a temporal network in which citation links are timestamped with the time of publication of the citing paper. Static ranking methods applied to dynamic and evolving networks have been found to exhibit important shortcomings [Liao et al., 2017], consequently a variety of dedicated methods have been put forward.

Several PageRank variants have been proposed which address the problem of the method's bias towards older nodes (see e.g. Nykl et al. [2014]; Fiala et al. [2015a]). Usually a decaying weighting scheme is introduced, in order to increasingly penalise older edges/nodes and favour recent ones (see *inter alia* Xing and Ghorbani [2004]; Ding [2011]; Yan and Ding [2011]; Fiala et al. [2015b]; Kong et al. [2015], for a review [Liao et al., 2017, 4.2]). Some other approaches exist. For example, Fiala [2012] used publication date information to weight the edges in author citation networks, weighted by considering co-authorship relations, while Jiang et al. [2016] proposed a cognitive interpretable ranking for articles in evolving networks, based on four steps: knowledge production, diffusion, accumulation and decay. The method also outputs time series which can be further analysed.

The dynamic method proposed in this paper borrows in part from methods developed in the context of the quantitative analysis of sport competitions. In particular, it is related to the Elo system [Elo, 1978; Langville and Meyer, 2012], a method coined by Arpad Elo to rank chess players and adopted by the World Chess Federation as the official rating system in chess. The method updates the rating of a player proportionally to the difference between the actual and expected performances of the player during a match. The expected performance is computed using a logistic function of the rating difference between the players of the match, similar to the one we used to define the citation rewards. Similar time-varying rating methods are typical in sport competitions, where season matches are distributed in time and there is the necessity of obtaining partial ratings for players or teams during the season; see for instance Motegi and Masuda [2012]; Cattelan et al. [2013]; Bozzo et al. [2017]. There are, however, some important differences between the dynamics of sport competitions and bibliometrics:

1. bibliometrics introduces an important asymmetry in the model: a citation, is, in most cases, a reward for the cited scholar, but it is not a penalty for the citing scholar;

2. bibliometrics is more flexible in the timing of the citations: while it is not possible for the same player to play twice at the same time, it is normal for a scholar to cite or be cited many times with the same timestamp;

3. the distribution of citations among scholars is typically very skewed, with few scholars collecting the majority of citations. On the other hand, in sport competitions, the number of played games among players is quite stable. For instance, in round-robin competitions (like soccer national leagues), all teams play the same number of games;

4. finally, a player cannot play against himself. On the other hand, self-citation is an established practice in bibliometrics.

Recent work applies the Elo method to the ranking of journals [Lehmann and Wohlrabe, 2017], and compare with established solutions such as the Source Normalized Impact per Publication (SNIP). The authors simulate a yearly round-robin ‘competition’ between each pair of different journals and consider the outcome of a match between journals a win for whoever has the higher SNIP score at that year, or a tie if the two journals have equal SNIP score. The journal ratings are updated every year using the Elo system. The authors suggest that the time line of competitions brings new information in the ranking of journals, which was not previously accounted for, and claim that the Elo ranking seems a promising alternative to already existing ranking approaches.

5 Conclusions

We presented TimeRank: a method for rating scholars which accounts for the dynamic nature of the scientific process. Our method updates the rating of a scholar when a citation is received, by considering the relative rating of the citing and cited scholars at the time of the citation. The method is demonstrably sensible to citation timing and consistent provided an appropriate parameter choice. With this system, the quantity, quality and timing of citations all contribute to the final rating. Interestingly, the rating for a scholar can be decomposed into a time series of citation rewards. The analysis of the temporal evolution of rewards can discriminate between similarly rated scholars that in fact have quite different citation reward histories.

We applied our method on the bibliomerics community finding that it behaves differently from alternatives such as total number of citations, PageRank and the Hirsch index, especially for the top rated scholars. The method specifically levels-out the distance between established researchers who were gradually recognized in their community, and rapidly rising stars who were able to accrue citations from highly rated scholars early on in their careers. Our proposed method can be best applied in situations when all quantity, quality and timing are relevant, for example by committees involved in the decision to hire early-stage researches, who might be interested in detecting rising stars (i.e. young scholars who received early recognition from higher-rated scholars).

Acknowledgments

Giovanni Colavizza worked on this research project during a research stay at the Centre for Science and Technology Studies (CWTS) at Leiden University. He would like to thank CWTS for providing access to its databases. Colavizza is supported by the Swiss National Fund with grant P1ELP2_168489. Franceschet acknowledges the University of Udine for a PRID 2017 ENCASE grant. The authors would finally like to thank Paolo Vidoni, Enrico Bozzo and Carla Piazza, as well as the anonymous referees, for their help and suggestions.

References

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

E. Bozzo, M. Franceschet, and P. Vidoni. The temporalized massey's method. *Journal of Quantitative Analysis in Sports*, 13(2):37–48, 2017.

S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *International World-Wide Web Conference*, 1998. Accessed March 1, 2010 from `http://ilpubs.stanford.edu:8090/361/`.

E. Caron and N. J. van Eck. Large scale author name disambiguation using rule-based scoring and clustering. In *Science and Technology Indicators Conference*, pages 79–86, 2014.

M. Cattelan, C. Varin, and D. Firth. Dynamic Bradley-Terry modelling of sports tournaments. *Journal of the Royal Statistical Society Series C Applied Statistics*, 62(1):135–150, 2013.

M. Charrad, N. Ghazzali, V. Boiteau, and A. Niknafs. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6):1–36, 2014. URL `http://www.jstatsoft.org/v61/i06/`.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `http://igraph.org`.

Y. Ding. Applying weighted PageRank to author citation networks. *Journal of the American Society for Information Science and Technology*, 62(2):236–245, 2011.

A. E. Elo. *The Rating of Chess Players, Past and Present.* Arco, New York, 1978.

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

D. Fiala, F. Rousselot, and K. Ježek. PageRank for bibliographic networks. *Scientometrics*, 76(1):135–158, 2008.

D. Fiala, L. Šubelj, S. Žitnik, and M. Bajec. Do PageRank-based author rankings outperform simple citation counts? *Journal of Informetrics*, 9(2):334–348, 2015a.

D. Fiala, G. Tutoky, P. Koncz, and J. Parali. Ageing of edges in collaboration networks and its effect on author rankings. *Acta Polytechnica Hungarica*, 12 (6):149–160, 2015b.

M. Franceschet. The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, 4(1):55–63, 2010.

D. Hicks, P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520(7548):429–431, 2015.

J. E. Hirsch. 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, 2005.

P. Holme. Modern temporal network theory: a colloquium. *European Physical Journal B*, 88:234, 2015.

P. Holme and J. Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.

X. Jiang, C. Gao, and R. Liang. Ranking Scientific Articles in a Dynamically Evolving Citation Network. In *Semantics, Knowledge and Grids (SKG), 2016 12th International Conference on*, pages 154–157. IEEE, 2016.

X. Kong, J. Zhou, J. Zhang, W. Wang, and F. Xia. TAPRank: A Time-Aware Author Ranking Method in Heterogeneous Networks. pages 242–246. IEEE, 2015.

A. N. Langville and C. D. Meyer. *Who's #1? The science of rating and ranking.* Princeton University Press, Princeton, NJ, 2012.

R. Lehmann and K. Wohlrabe. Who is the 'Journal Grand Master'? A new ranking based on the Elo rating system. *Journal of Informetrics*, 11(3):800 – 809, 2017.

L. Leydesdorff, P. Zhou, and L. Bornmann. How can journal impact factors be normalized across fields of science? An assessment in terms of percentile ranks and fractional counts. *Journal of the American Society for Information Science and Technology*, 64(1):96–107, 2013.

H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou. Ranking in evolving complex networks. *Physics Reports*, 2017.

T. Marchant. An axiomatic characterization of the ranking based on the h-index and some other bibliometric rankings of authors. *Scientometrics*, 80 (2):325–342, 2009a.

T. Marchant. Score-based bibliometric rankings of authors. *Journal of the American Society for Information Science and Technology*, 60(6):1132–1137, 2009b.

S. Motegi and N. Masuda. A network-based dynamical ranking system for competitive sports. *Scientific Reports*, 2(904), 2012.

V. Nicosia, J. Tang, C. Mascolo, M. Musolesi, G. Russo, and V. Latora. Graph metrics for temporal networks. In P. Holme and J. Saramäki, editors, *Temporal networks*, Understanding Complex Systems, pages 15–40. Springer-Verlag Berlin Heidelberg, 2013.

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

M. Olensky, M. Schmidt, and N. J. van Eck. Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the Web of Science. *Journal of the Association for Information Science and Technology*, 67(10):2550–2564, 2016.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1999. URL `http://ilpubs.stanford.edu:8090/422/`. Previous number = SIDL-WP-1999-0120.

A. Perianes-Rodriguez, L. Waltman, and N. J. van Eck. Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4):1178–1195, 2016.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

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

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

L. Waltman and N. J. Van Eck. A taxonomy of bibliometric performance indicators based on the property of consistency. Technical Report ERS-2009-014-LIS, Erasmus Research Institute of Management, 2009. URL `http://repub.eur.nl/pub/15182/ERS-2009-014-LIS.pdf`.

L. Waltman and N. J. van Eck. The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2):406–415, 2012.

L. Waltman and E. Yan. PageRank-related methods for analyzing citation networks. In *Measuring scholarly impact*, pages 83–100. Springer, 2014.

J. D. West, M. C. Jensen, R. J. Dandrea, G. J. Gordon, and C. T. Bergstrom. Author-level Eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community.

Journal of the American Society for Information Science and Technology, 64 (4):787–801, 2013.

L. Wildgaard, J. W. Schneider, and B. Larsen. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, 101(1):125–158, 2014.

W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314. IEEE, 2004.

E. Yan and Y. Ding. Discovering author impact: A PageRank perspective. *Information Processing & Management*, 47(1):125–134, 2011.

E. Yan, Y. Ding, and C. R. Sugimoto. P-Rank: An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology*, 62(3):467–477, 2011.

C. Zhang, C. Liu, L. Yu, Z.-K. Zhang, and T. Zhou. Identifying the Academic Rising Stars. *arXiv preprint arXiv:1606.05752*, 2016. URL `https://arxiv.org/abs/1606.05752`.

Y.-B. Zhou, L. Lü, and M. Li. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics*, 14(3):033033, 2012.