systems

MDPI

Article

One-Bit In, Two-Bit Out: Network-Based Metrics of Papers Can Be Largely Improved by Including Only the External Citation Counts without the Citation Relations

Jianlin Zhou [1], Zhesi Shen [2] and Jinshan Wu [1,*]

[1] School of Systems Science, Beijing Normal University, Beijing 100875, China; jianlinzhou@bnu.edu.cn
[2] National Science Library, Chinese Academy of Sciences, Beijing 100190, China; shenzhs@mail.las.ac.cn
* Correspondence: jinshanw@bnu.edu.cn

Abstract: Many ranking algorithms and metrics have been proposed to identify high-impact papers. Both the direct citation counts and the network-based PageRank-like algorithms are commonly used. Ideally, the more complete the data on the citation network, the more informative the ranking. However, obtaining more data on citation relations is often costly and challenging. In some cases, obtaining the citation counts can be relatively simple. In this paper, we look into using the additional citation counts but without additional citation relations to form more informative metrics for identifying high-impact papers. As an example, we propose enhancing the original PageRank algorithm by combining the local citation network with the additional citation counts from a more complete data source. We apply this enhanced method to American Physical Society (APS) papers to verify its effectiveness. The results indicate that the proposed ranking algorithm is robust against missing data and can improve the identification of high-quality papers. This shows that it is possible to enhance the effectiveness of a network-based metric calculated on a relatively small citation network by including only the additional data of the citation counts, without the additional citation relations.

Keywords: citation network; complete data; PageRank; citation count

check for updates

Citation: Zhou, J.; Shen, Z.; Wu, J. One-Bit In, Two-Bit Out: Network-Based Metrics of Papers Can Be Largely Improved by Including Only the External Citation Counts without the Citation Relations. *Systems* **2024**, *12*, 377. https://doi.org/10.3390/systems12090377

Academic Editor: Gianfranco Minati

Received: 22 July 2024
Revised: 11 September 2024
Accepted: 13 September 2024
Published: 17 September 2024

1. Introduction

Quantifying the impact of papers is an important research topic in scientometrics. As the number of publications grows exponentially [1], it has become increasingly important. Based on the measured impact of papers, one can design algorithms to automatically recommend high-quality papers to scientists, which can help them to understand the frontiers of science and create knowledge [2,3]. However, quantifying the impact of a paper is not an easy task because some mechanisms, such as preferential attachment [4,5], aging [6], and fitness [7,8], all play critical roles in the citation dynamics of a paper [9]. As such, one often wants to focus on, or hopefully look into, ultimately measuring the scientific value or creativity of papers. Until now, many evaluation indicators have been proposed to solve this problem [10–15]. They all measure proxies of, but do not measure directly, the scientific value or creativity of papers.

In this work, we are not aiming to solve this problem of defining new metrics to measure papers' scientific value or creativity directly. Instead, we want to focus on a very technical problem: the issue of incomplete data [16,17]. When developing and applying the impact metrics of papers, we often work with a particular dataset. For example, one may use the American Physical Society (APS) papers and the citation relation among the papers within the APS journals, papers in Web of Science (WoS), and their citation relation within the WoS journal coverage, or Scopus, Dimensions, and so on. We have to note that even the large datasets, such as WoS, Scopus, and Dimensions, are still incomplete, meaning that there are missing papers and, thus, missing citation relations, for example, from journals not covered by the dataset. Let us use APS data as an example. On the one hand, APS papers can be more or less regarded as a core set of all physics papers. Let us even assume that, for now, we only care about APS papers and want to rank them and then recommend them to readers according to the rank. Thus, ranking other physics papers is irrelevant for now. On the other hand, those APS papers are not cited by only APS papers. Therefore, even if we do not want to rank other papers, the citations from other papers might also help rank the APS papers. Now, the question becomes how we can use the citations from the other papers to rank the APS papers. Of course, we can include the other papers in the citation network and apply the analysis to calculate the corresponding ranking metrics again. However, for that, we will need to incorporate the complete citation relationships among the other papers and their citations to the APS papers. Gaining access to the citation relationships among other papers is very costly, and the expanded network may significantly increase in size. It will be very cost-effective if we can improve the metrics using the same citation network, but with additional data only on the citation counts from the other papers to the APS papers.

If we can show that including the additional citation counts but not the citation relations can improve one of the current network-based metrics (here, the original PageRank ranking) calculated on a relatively small citation network, then the same can be performed for all other network-based metrics, including those that have improved upon the original PageRank algorithm to deal with, for example, the aging effect [18] and some other nonlinear considerations [19,20]. Why do we choose the original PageRank as an example for this investigation, and can we use the counts of direct citations instead? The former is a network-based metric, which considers both the direct and indirect citation relations, while the latter uses only the direct citations. For the latter, clearly including the additional citation counts will improve the metric, but there is no additional gain. The amount of additional information that is included determines how the amount of gain. However, with the former, the additional citation counts in the enhanced calculation will be propagated over the current citation network. Therefore, it is possible that including only the citation counts in the metric brings more than the citation counts themselves.

More generally speaking, there is a related question of how complete will be enough for scientometric data. Are APS data complete enough? Are large data such as WoS, Scopus, or Dimensions complete enough? Are there ways to measure the completeness of the data, and are there ways to improve the metrics calculated on incomplete data with minimum cost? In this work, we focus on the last question, and we believe that a study of this question can help find a criterion of data completeness. For example, one may compare the gained information among the metrics calculated on the smaller dataset, the larger dataset, and the minimum cost approach. If expanding the dataset will lead to linear gain, then the dataset is not that complete, while a sub-linear marginal or diminishing gain might imply that the dataset is sufficiently complete. We will look into this line of investigation in future studies. For that, we need to have ways to extract gained information with the minimum-cost approach. This is the task in the current investigation.

In this paper, we introduce the additional citation counts to the PageRank algorithm and propose an external citation enhanced PageRank (exPRank) algorithm. We apply the exPRank algorithm on the APS citation network with additional citation counts of the same papers from WoS and apply other ranking algorithms on the APS citation network. We first check the effect of network incompleteness on the evaluation algorithms and then compare these ranking metrics, especially by looking at their ability to identify widely recognized high-quality papers, including the Nobel prize-winning papers and milestone papers selected by peers. The experimental results show that the exPRank algorithm is more robust and can better identify high-quality papers than other algorithms in the case of incomplete scientometrics data.

2. Data and Methods

2.1. Data in the Study

This study uses the APS data as the local citation network data and the Web of Science as the external citation count data. The American Physical Society provides the APS dataset, which includes 9 physics journals: Physical Review A, B, C, D, E, Letters, Series I & II, Special Topics, and Reviews of Modern Physics, from 1893 to 2010. It contains 482,577 papers with 5,016,422 citations[1]. In addition, the dataset also provides the DOI, title, author names, and affiliations of each APS paper. We retrieved their citation counts and the number of references in the Web of Science Core Collection database for each APS paper based on their DOIs in December 2019. We obtained 476,848 APS papers' citation counts and the number of references within WoS. In addition, we also found that the citation count of each paper within the APS dataset did not exceed its citation count in the Web of Science dataset.

2.2. External Citation Enhanced PageRank

The standard PageRank algorithm was originally proposed to evaluate the importance of web pages [21]. Later, it was extended to quantify the scientific impact of papers based on the citation network. Given a citation network with N nodes, PageRank is defined as

$$PR^i = \alpha \sum_{j=1}^{N} \left[\frac{A_j^i}{\sum_k A_j^k} \right] PR^j + \frac{1-\alpha}{N} \tag{1}$$

where $A_j^i = 1$ if paper j cites paper i and $A_j^i = 0$ otherwise. $\alpha \in [0,1]$ is a tuning parameter, whose value will affect the scores of nodes. The parameter α is usually set to 0.85.

Equation (1) can be written in the vector equation form as follows:

$$\boldsymbol{PR} = \alpha L \boldsymbol{PR} + (1-\alpha)\frac{1}{N}\boldsymbol{e}. \tag{2}$$

where $L = \left(\frac{A_j^i}{\sum_k A_j^k} \right)_{N \times N}$ is a transition probability matrix and $\boldsymbol{e}$ is a column vector having each component equal to 1. This equation can either be solved by matrix inverse as

$$\boldsymbol{PR} = (I - \alpha L)^{-1}\frac{1-\alpha}{N}\boldsymbol{e}. \tag{3}$$

or more practically by iterations such as

$$\boldsymbol{PR}(t) = \alpha L \boldsymbol{PR}(t-1) + (1-\alpha)\frac{1}{N}\boldsymbol{e}. \tag{4}$$

Finally, we rank the papers by sorting their final PageRank scores in descending order.

To introduce the total citation counts, C_{ex}^i, from the large-scale database into PageRank, we proposed the external citation enhanced PageRank. First, we added two virtual nodes $N+1$ and $N+2$ into the citation network, where the node $N+1$ will cite some nodes in the original network and the node $N+2$ will be cited by some nodes in the original citation network. Then, for any node i in the original network, we modified the citation matrix A by

$$A_{N+1}^i = \theta\left(C_{\text{ex}}^i - C_{\text{in}}^i\right)\left(C_{\text{ex}}^i - C_{\text{in}}^i\right). \tag{5}$$

$$A_i^{N+1} = 0. \tag{6}$$

$$A_i^{N+2} = \theta\left(C_i^{\text{ex}} - C_i^{\text{in}}\right)\left(C_i^{\text{ex}} - C_i^{\text{in}}\right). \tag{7}$$

$$A_{N+2}^i = 0. \tag{8}$$

where $\theta(x)$ is the step function; $\theta(x) = 1$ if $x \geq 0$, and $\theta(x) = 0$ otherwise. $C^i_{\mathrm{in}} = \sum_{j=1}^{N} A^i_j$ is the citation count of node i within the original citation network, $C_i^{\mathrm{in}} = \sum_{j=1}^{N} A_i^j$ is the number of references of node i within the original citation network, and C_i^{ex} represents the number of references of node i within the large-scale database.

Plug this newly extended matrix A and vector $\boldsymbol{PR}$, which, since now being extended to objects with dimension $d = N + 2$, is called exPRank score and denoted as $\boldsymbol{exPR}$, back into Equation (2) and keep only those $exPR^i$ for $i \leq N$. Then, we obtain the recursive equation of exPRank:

$$exPR^i = \alpha \sum_{j=1}^{N} L^i_j exPR^j - \alpha \sum_{j=1}^{N} \sum_{k=1}^{N} L^k_j exPR^j Q^i + \frac{2 + N\alpha}{N+2} Q^i + \frac{1-\alpha}{N+2}, \tag{9}$$

where

$$Q^i = \frac{C^i_{\mathrm{ex}} - C^i_{\mathrm{in}}}{\sum_j \left(C^j_{\mathrm{ex}} - C^j_{\mathrm{in}} \right)}. \tag{10}$$

In matrix form, it can be written as

$$\boldsymbol{exPR} = \alpha \left(I - \boldsymbol{Q}\boldsymbol{e}^{\mathrm{T}} \right) L\boldsymbol{exPR} + \frac{2 + N\alpha}{N+2} \boldsymbol{Q} + \frac{1-\alpha}{N+2} \boldsymbol{e}. \tag{11}$$

Equation (11) is equivalent to

$$\begin{aligned} \boldsymbol{exPR} &= \left(I - \alpha \left(I - \boldsymbol{Q}\boldsymbol{e}^{\mathrm{T}} \right) L \right)^{-1} \left(\frac{2 + N\alpha}{N+2} \boldsymbol{Q} + \frac{1-\alpha}{N+2} \boldsymbol{e} \right) \\ &= \frac{1}{N+2} \sum_{n=0}^{\infty} \alpha^n \left(\left(I - \boldsymbol{Q}\boldsymbol{e}^{\mathrm{T}} \right) L \right)^n ((2 + N\alpha)\boldsymbol{Q} + (1-\alpha)\boldsymbol{e}), \end{aligned} \tag{12}$$

from which we can clearly see how the number of external additional citations $\boldsymbol{Q}$ is propagated into the system, order by order, via the various powers of L.

It is exactly this propagation, or, in mathematical terms, L^n, that makes it possible that the gained amount of information output can be more than the input data, and we need to test it to see whether or not, in practice, on some paper-level indicators, one-bit input can lead to two-bit (or more accurate more than one bit) output. Using direct citation count C^i as an example, then there is clearly no such additional bit of gain; since the $\Delta C^i = C^i_{\mathrm{ex}}$, there is one-bit output for one-bit input. Will a network-based indicator make it possible to have additional gain? In this paper, we apply the exPRank to the APS local citation network with additional citation counts and apply the original PageRank to the APS citation network. We then compare the revealed information from these metrics to check the amount of gained information.

2.3. Other Compared Metrics

This paper compares the external citation-enhanced PageRank against four other indicators. The first indicator is the standard PageRank. The second indicator is PrestigeRank (PrRank), which is proposed by Su et al. [22] to reduce the effect of missing data in the citation database on the results of PageRank. This algorithm also introduces a virtual node that connects all the nodes. The virtual node is supposed to represent those references not included in the citation database, and it receives all citations that are from papers in the database. The formula of the PrestigeRank algorithm is expressed as

$$\boldsymbol{\pi}(t) = \left[\alpha L + \frac{\boldsymbol{e}}{N} \left(\alpha \boldsymbol{a}^{\mathrm{T}} + (1-\alpha) \boldsymbol{e}^{\mathrm{T}} \right) \right] \boldsymbol{\pi}(t-1) \tag{13}$$

where $L = \left(L_{ij}\right)_{(N+1)\times(N+1)}$. When $i \leq N$ and $j \leq N$, $L_{ij} = \frac{A_j^i}{C_j^{\mathrm{ex}}}$. $L_{(N+1)j} = 1 - \sum_{i=1}^{N} L_{ij}$, $L_{i(N+1)} = \frac{C_{\mathrm{in}}^i}{\sum_{i=1}^{N} C_{\mathrm{in}}^i + \sum_{i=1}^{N}\left(C_i^{\mathrm{ex}} - C_i^{\mathrm{in}}\right)}$. $\boldsymbol{a}$ is the binary dangling node vector.

The remaining indicators are two types of citation counts that are among the simplest ways to measure the impact of a paper. One is calculated based on the local citation network constructed by the APS dataset, and we use $C_{\mathrm{APS}}(i)$ (or C_{in}^i in Equation (5)) to represent the citation count within the APS dataset for paper i. Another is the citation count $C_{\mathrm{WoS}}(i)$ (or C_{ex}^i in Equation (5)) retrieved from the Web of Science, representing the global citation count.

3. Results

3.1. The Analysis of the Incompleteness of the APS Dataset

Based on the number of citations of APS papers in the APS and WoS datasets, we first examine the incompleteness of the APS dataset. The total number of citations of APS papers in the APS dataset is 4,945,687, while the total number of citations of these papers in the WoS dataset is 24,018,751, which indicates that the APS dataset is a local dataset and it has many missing data points. In Figure 1, we show the citation count distribution of APS papers within APS (C_{APS}) and Web of Science (C_{WoS}). One can see that they both approximately follow the power-law distribution, and the distribution of C_{WoS} is much broader than the distribution of C_{APS} in Figure 1A. We also calculate the cumulative distribution of ($C_{\mathrm{WoS}} - C_{\mathrm{APS}}$) and the result is shown in Figure 1B. We observe that only 5.18% APS papers have the same number of citations in the APS and WoS datasets, and 17.96% papers have a value of ($C_{\mathrm{WoS}} - C_{\mathrm{APS}}$) exceeding 50. Therefore, there is an obvious difference in the number of citations of papers between large and small citation databases, which will affect the evaluation of papers in small databases. Now, the question becomes whether or not we can make efficient and low-cost use of this difference, via the exPRank defined in Equation (9), to improve the performance of the PageRank algorithm in ranking papers in a small citation database.

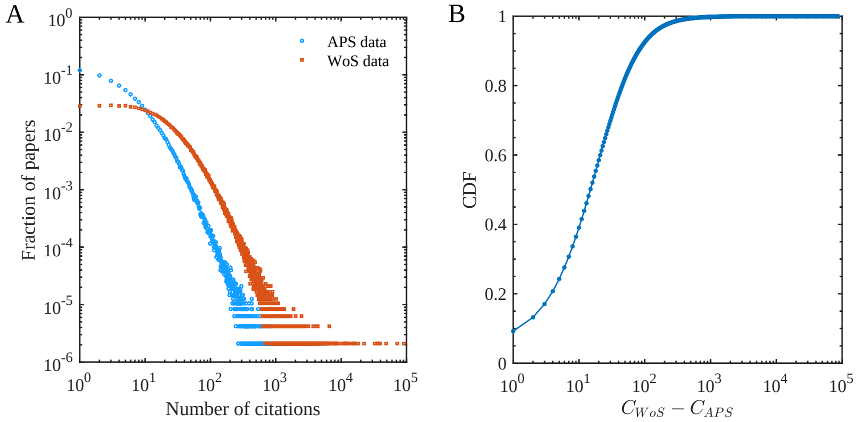

Figure 1. (**A**) The distribution of the C_{APS} and C_{WoS} of papers. (**B**) The cumulative distribution of the ($C_{\mathrm{WoS}} - C_{\mathrm{APS}}$) of papers.

3.2. The Effect of an Incomplete Citation Network on Evaluation Algorithms

Assuming that the APS citation network constructed based on APS data is a complete network (that is, for each paper in the citation network, all its citing papers and cited papers can be found in this citation network), then many incomplete networks can be obtained by deleting various different proportions of links in the above complete network. We test the effect of network incompleteness on the evaluation algorithms. We first apply the PageRank algorithm to this complete citation network, obtain the PageRank score for each paper, and calculate the number of citations of papers in the citation network. Then, we consider the number of citations in the complete citation network as additional citation counts and apply the exPRank and PageRank algorithms on these incomplete citation networks, to obtain the citation counts of papers in these incomplete citation networks. We calculate the Spearman correlation coefficients between PageRank scores on the complete citation network and exPRank scores on the incomplete citation networks, as well as PageRank scores on the incomplete citation networks, for which strong correlation indicates higher robustness of the evaluation algorithm against network incompleteness. We also calculate the Spearman correlation coefficients between citation counts within the complete citation network and citation counts within the incomplete citation networks. The results are shown in Figure 2. We can find that the exPRank algorithm has higher robustness against network incompleteness than the PageRank algorithm. By continuously removing links from the original citation network, we can observe that the correlation coefficients between PageRank scores on the initial citation network and PageRank scores on the incomplete citation networks decrease first slowly and then rapidly. The metric of citation counts also shows similar results to PageRank. However, the correlation coefficients between PageRank scores on the initial citation network and exPRank scores on the incomplete citation networks have been decreasing relatively slowly, with correlation coefficients basically above 0.9. This analysis shows that the exPRank algorithm can bring much information just by using the additional citation counts.

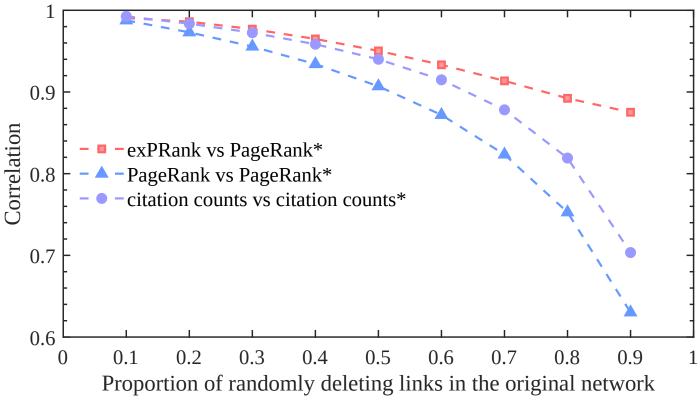

Figure 2. The correlation between the PageRank algorithm on the complete citation network and exPRank algorithm on the incomplete citation network, as well as the PageRank algorithm on the complete citation network. “PageRank*” represents the results of PageRank algorithm applied to the complete citation network. “citation counts*” represents the results of citation counts of papers in the complete citation network. Each point in the figure is averaged over 20 realizations of the networks by randomly deleting links in the complete citation network.

3.3. Correlation between Evaluation Algorithms

We apply the exPRank and other algorithms to the citation network constructed based on APS data. We then calculate the Spearman correlation coefficient between the scores of these algorithms, and the results are presented in Table 1. One can see that there is a positive correlation in the exPRank, PageRank, PrestigeRank, citation count based on APS data, and citation count based on WoS data. The exPRank shows a higher correlation with C_{WoS}, but a relatively low correlation with the PageRank, PrestigeRank, and C_{APS}. The PageRank shows a higher correlation with PrestigeRank and C_{APS}, but a relatively low correlation with the exPRank and C_{WoS}. These can show that the exPRank is different from PageRank. The Spearman correlation coefficient between C_{APS} and C_{WoS} is only 0.7541, which also indicates that there are obvious differences between these two indicators in terms of numerical value and ranking.

We further investigate the overlap of the top-ranking papers identified by these indicators. As can be seen from Table 2, among these pairs of indices, the pairs of exPRank–C_{WoS} and PrestigeRank–C_{APS} have a higher overlap ratio, and their overlap ratios are both more than 0.6. The pairs of PageRank–C_{WoS} and PageRank–C_{APS} have a lower overlap rate, which does not exceed 0.25. In addition, although PageRank has a high correlation with PrestigeRank and C_{APS}, we find that the overlap ratio of the top 1% of papers between PageRank and any other indicator is not high.

Table 1. The Spearman rank correlation coefficients of different ranking algorithms.

Algorithms	exPRank	PageRank	PrestigeRank	C_{APS}	C_{WoS}
exPRank	1	0.7223 ($p < 0.01$)	0.7717 ($p < 0.01$)	0.7766 ($p < 0.01$)	0.9750 ($p < 0.01$)
PageRank	0.7223 ($p < 0.01$)	1	0.9366 ($p < 0.01$)	0.8905 ($p < 0.01$)	0.6448 ($p < 0.01$)
PrestigeRank	0.7717 ($p < 0.01$)	0.9366 ($p < 0.01$)	1	0.9844 ($p < 0.01$)	0.7178 ($p < 0.01$)
C_{APS}	0.7766 ($p < 0.01$)	0.8905 ($p < 0.01$)	0.9844 ($p < 0.01$)	1	0.7541 ($p < 0.01$)
C_{WoS}	0.9750 ($p < 0.01$)	0.6448 ($p < 0.01$)	0.7178 ($p < 0.01$)	0.7541 ($p < 0.01$)	1

Table 2. The overlap ratio of the top 1% of papers identified by different ranking algorithms.

Algorithms	exPRank	PageRank	PrestigeRank	C_{APS}	C_{WoS}
exPRank	1	0.2905	0.5104	0.4827	0.6376
PageRank	0.2905	1	0.3593	0.2366	0.1912
PrestigeRank	0.5104	0.3593	1	0.6370	0.3587
C_{APS}	0.4827	0.2366	0.6370	1	0.4501
C_{WoS}	0.6376	0.1912	0.3587	0.4501	1

Table 3 lists the top-10 papers ranked by the exPRank and their corresponding rankings under different algorithms. We can observe that the rankings of these top-10 papers in Prestigerank, C_{APS}, and C_{WoS} are relatively high, but their ranking in PageRank is relatively low. Overall, the ranking results of these 10 papers under the exPRank are very close to those of C_{WoS}, but it is not that close to those of PageRank. For example, the rank of the paper (10.1103/PhysRevLett.77.3865) based on the exPRank is 2, but its rank in PageRank is 3073. In addition, we can clearly see the effect of the scale of the citation database on the evaluation of papers. For instance, the C_{APS} values of some papers (e.g., 10.1103/PhysRevB.37.785, 10.1103/PhysRevA.38.3098) are relatively small, but their C_{WoS} values are relatively large. Paper 10.1103/PhysRevB.37.785 was selected as one of Physical Review B's 50th Anniversary Milestones in 2020. It has been cited 72,092 times in the WoS dataset, but it has been cited only 656 times in the APS database. The citation count of Paper 10.1103/PhysRevA.38.3098 is only 728 times in the APS dataset but 38,670 times in the WoS dataset. It was selected as one of Physical Review A's 50th Anniversary Milestones in 2020.

Table 3. The top-10 papers selected by exPRank and their corresponding rankings in other algorithms.

Paper	Rank					Citation Counts	
	exPRank	PageRank	Prestigerank	C_{APS}	C_{WoS}	APS	WoS
PhysRevLett.77.3865	1	273	5	3	1	3690	91,229
PhysRevB.37.785	2	3073	365	107	2	656	72,092
PhysRev.140.A1133	3	8	1	1	4	5560	39,460
PhysRevB.54.11169	4	567	13	6	3	2818	48,754
PhysRev.136.B864	5	13	2	2	9	4399	32,749
PhysRevA.38.3098	6	2106	193	87	5	728	38,670
PhysRevB.13.5188	7	257	7	5	8	2843	35,842
PhysRevB.50.17953	8	1391	38	10	6	1801	36,268
PhysRevB.59.1758	9	1517	43	11	7	1784	35,893
PhysRevB.23.5048	10	83	3	4	14	3325	15,436

3.4. Identifying High-Quality Papers

Finally, we select some high-quality papers recognized by scientists as benchmarks to evaluate the performance of the ranking algorithms. Our benchmarks are the 70 Nobel prize-winning papers in Physics, 87 Physical Review Letters (PRL) milestone papers, 23 Physical Review A (PRA) milestone papers, 47 Physical Review B (PRB) milestone papers, 23 Physical Review E (PRE) milestone papers, and 74 selected papers for celebrating 125 years of the Physical Review journals. The Nobel prize-winning physics papers reflect the main contributions or work of the Nobel laureates in Physics. We collected these Nobel prize-winning physics papers based on some of the influential literature [23,24]. The milestone papers in the other five datasets were chosen by the editors of APS journals to be regarded as the most significant contributions to Physics from each journal. These five datasets can be obtained from these links[2].

Taking the Nobel prize-winning physics papers and PRL milestone papers as examples, we compare the rank of these recognized high-impact papers on exPRank and other ranking algorithms in Figure 3. We can first find that, compared with ordinary papers, these high-quality papers are usually ranked higher with respect to any pairs of those four quantities, i.e., more or less concentrated towards the top-right corner in all four figures. Second, more or less, it is more in the right direction than in the up direction, especially in Figure 3A–C. This implies that the exPRank algorithm is more informative in identifying those high-quality papers. Lastly, those high-quality papers are more or less concentrated along the diagonal line in Figure 3D. This means that, for those high-quality papers, their exPRank scores are very similar to C_{WoS}.

To make the difference among the five indicators more visible, for each dataset, we calculate the average rankings of high-quality papers under the scores of the different indicators, and the results are shown in Figure 4. The lower the average ranking value of high-quality papers, the higher their ranking. Overall, the exPRank performs the best in identifying high-quality papers among these indicators. PageRank and C_{WoS} perform second to exprank, while C_{APS} performs worst. It should be pointed out that PageRank has a better performance in identifying Nobel Prize-winning papers and PRL milestone papers. C_{WoS} performs best in identifying PRA, PRB, and PRE milestone papers.

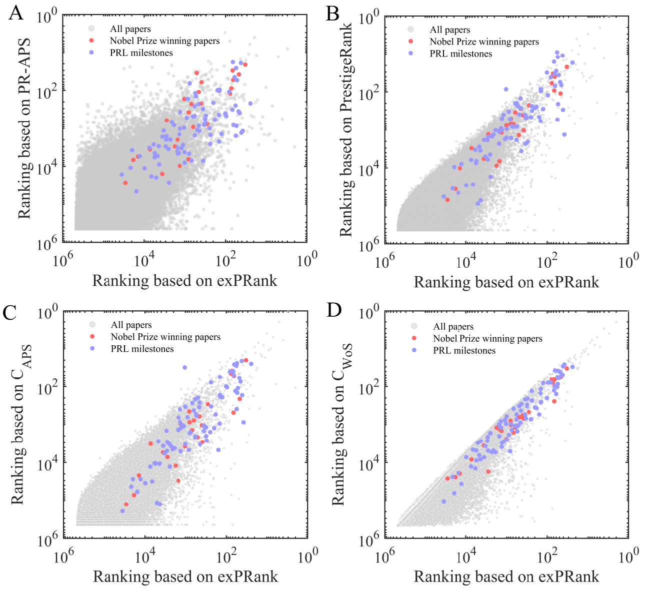

Figure 3. The scatter plots of ranking results for exPRank versus other ranking algorithms. (**A**) Comparison of the ranking of all papers under exPRank and PageRank. (**B**) Comparison of the ranking of all papers under exPRank and PrestigeRank. (**C**) Comparison of the ranking of all papers under exPRank and C_{APS}. (**D**) Comparison of the ranking of all papers under exPRank and C_{WoS}. Nobel prize-winning physics papers and PRL milestone papers are typical high-quality papers, and their rankings under different algorithms are represented by red and purple dots.

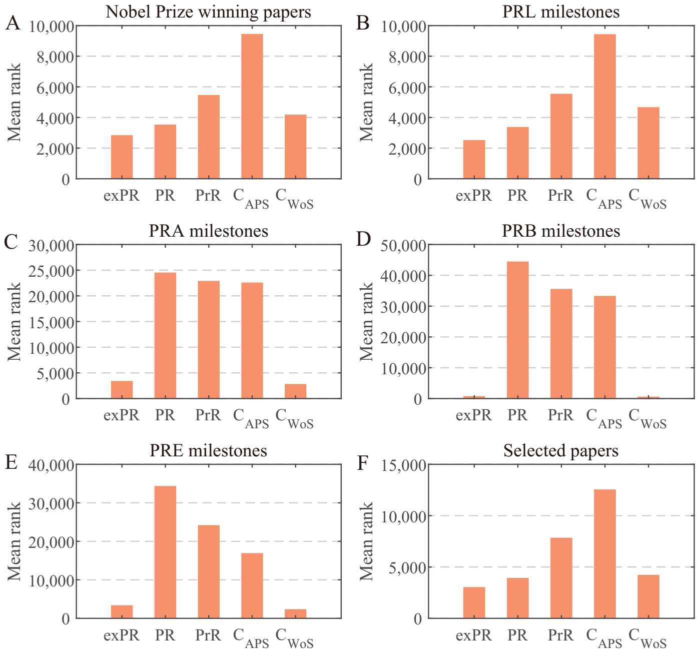

Figure 4. The mean rank of (**A**) 70 Nobel prize-winning papers in Physics, (**B**) 87 Physical Review Letters milestone papers, (**C**) 23 Physical Review A milestone papers, (**D**) 47 Physical Review B milestone papers, (**E**) 23 Physical Review E milestone papers, and (**F**) 74 selected papers for celebrating 125 years of the Physical Review journals in different ranking algorithms. "exPR" represents the external citation enhanced PageRank, "PR" represents PageRank, and "PrR" represents Prestigerank.

The fact that, overall, the exPRank performs better than C_{WoS} shows that the gained information from exPRank with additional data on citation counts but without citation networks is more than that from C_{WoS}, which leads to exactly one bit of information out with each bit of additional input data.

4. Conclusions and Discussion

In this paper, we have shown that, by including only the additional citation counts but not the additional citation relations from a large dataset (Web of Science data) into a small dataset, the performance of a metric of impact, particularly, the PageRank score, on a relatively smaller dataset (the American Physical Society data) can be significantly improved. This same approach of utilizing the less expensive additional citation count data, instead of the more costly citation relationship data, can also be applied to other metrics calculated on smaller datasets.

We want to point out that the additional data on citation counts can only modify the direct citation counts of papers in the small dataset. Thus, at most, the gained informativeness using the additional citation counts can be as high as the complete citation counts, which in our case is the C_{WoS}. However, we can see from our results in Figure 4 that the gained improvement on exPRank is more than that of C_{WoS}. This is because exPRank is a network-based indicator.

We should also note that, although exPRank seems more informative than the other four indicators, much of the gain is due to the effectiveness of C_{WoS}. In fact, exPRank can be viewed as a hybrid of PageRank and C_{WoS}, and it takes and combines the better of PageRank and C_{WoS}. While this shows that network-based indicators can turn one bit of additional input data into more than one bit of output gain in indicators and their performances, there might be an even better way to do it. In addition, there are also cases where exPRank is even slightly worse than C_{WoS}, and, in those cases, the performance of PageRank is much worse than that of C_{WoS}. This might be used to indicate the severe degree of the incompleteness of the small dataset.

More generally, since we now have ways to improve metric calculated on small datasets with a not-so-costly approach, we should be able to come back to the question of measuring the degree of the completeness of data on citation relations (or any network, for that matter). We can compare the effectiveness of the metric calculated on the smaller dataset, the one calculated on the larger dataset, and the one calculated on the smaller dataset but enhanced by cheap data from the larger dataset. The lower the gain from metrics that use additional data on the small dataset, the more complete the original dataset. Additionally, the function of each edge in the network is not the same, and some edges play a more important role than others in defining the structures and functions of the network. Thus, some simple statistics, such as link density, will not be a good measure of the completeness of the data. The present work might help investigate such a measure and find ways to improve upon it in the future.

Author Contributions: Conceptualization, J.W.; methodology, J.Z., Z.S. and J.W.; software, J.Z.; validation, J.Z., Z.S. and J.W.; data curation, J.Z.; writing—original draft preparation, J.Z., Z.S. and J.W.; writing—review and editing, J.Z., Z.S. and J.W.; supervision, Z.S. and J.W.; funding acquisition, J.Z. and J.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) (Grant No. 72074030) and the MOE (Ministry of Education in China) Liberal Arts and Social Sciences Foundation (Grant No. 20YJC870015), the Fundamental Research Funds for the Central Universities.

Data Availability Statement: The data are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

[1] These data can be obtained by submitting a request via https://journals.aps.org/datasets (accessed on 21 November 2021).

[2]
- https://journals.aps.org/prl/50years/milestones (accessed on 21 November 2021);
- https://journals.aps.org/pra/50th (accessed on 21 November 2021);
- https://journals.aps.org/prb/50th (accessed on 21 November 2021);
- https://journals.aps.org/pre/collections/pre-milestones (accessed on 21 November 2021);
- https://journals.aps.org/125years (accessed on 21 November 2021).

References

1. Fortunato, S.; Bergstrom, C.T.; Börner, K.; Evans, J.A.; Helbing, D.; Milojević, S.; Petersen, A.M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; et al. Science of science. *Science* **2018**, *359*, eaao0185. [CrossRef] [PubMed]
2. Ali, Z.; Qi, G.; Muhammad, K.; Ali, B.; Abro, W.A. Paper recommendation based on heterogeneous network embedding. *Knowl.-Based Syst.* **2020**, *210*, 106438. [CrossRef]
3. Ke, Q.; Gates, A.J.; Barabási, A.L. A network-based normalized impact measure reveals successful periods of scientific discovery across disciplines. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2309378120. [CrossRef] [PubMed]
4. Barabási, A.L.; Albert, R. Emergence of scaling in random networks. *Science* **1999**, *286*, 509–512. [CrossRef]
5. Jeong, H.; Néda, Z.; Barabási, A.L. Measuring preferential attachment in evolving networks. *Europhys. Lett.* **2003**, *61*, 567–572. [CrossRef]
6. Medo, M.; Cimini, G.; Gualdi, S. Temporal effects in the growth of networks. *Phys. Rev. Lett.* **2011**, *107*, 238701. [CrossRef]
7. Bianconi, G.; Barabási, A.L. Competition and multiscaling in evolving networks. *Europhys. Lett.* **2001**, *54*, 436–442. [CrossRef]
8. Caldarelli, G.; Capocci, A.; De Los Rios, P.; Munoz, M.A. Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **2002**, *89*, 258702. [CrossRef]
9. Wang, D.; Song, C.; Barabási, A.L. Quantifying long-term scientific impact. *Science* **2013**, *342*, 127–132. [CrossRef]
10. Zeng, A.; Shen, Z.; Zhou, J.; Wu, J.; Fan, Y.; Wang, Y.; Stanley, H.E. The science of science: From the perspective of complex systems. *Phys. Rep.* **2017**, *714*, 1–73. [CrossRef]
11. Bai, X.; Pan, H.; Hou, J.; Guo, T.; Lee, I.; Xia, F. Quantifying success in science: An overview. *IEEE Access* **2020**, *8*, 123200–123214. [CrossRef]
12. Zhang, F.; Wu, S. Measuring academic entities' impact by content-based citation analysis in a heterogeneous academic network. *Scientometrics* **2021**, *126*, 7197–7222. [CrossRef]
13. Leydesdorff, L.; Tekles, A.; Bornmann, L. A proposal to revise the disruption indicator. *Prof. Inf.* **2021**, *30*, e300121.
14. Leydesdorff, L.; Bornmann, L. Disruption indices and their calculation using web-of-science data: Indicators of historical developments or evolutionary dynamics? *J. Inf.* **2021**, *15*, 101219.

15. Yang, A.J.; Gong, H.; Wang, Y.; Zhang, C.; Deng, S. Rescaling the disruption index reveals the universality of disruption distributions in science. *Scientometrics* **2024**, *129*, 561–580. [CrossRef]
16. Lin, J.; Yu, Y.; Song, J.; Shi, X. Detecting and analyzing missing citations to published scientific entities. *Scientometrics* **2022**, *127*, 2395–2412. [CrossRef]
17. Delgado-Quirós, L.; Aguillo, I.F.; Martín-Martín, A.; López-Cózar, E.D.; Orduña-Malea, E.; Ortega, J.L. Why are these publications missing? Uncovering the reasons behind the exclusion of documents in free-access scholarly databases. *J. Assoc. Inf. Sci. Technol.* **2024**, *75*, 43–58. [CrossRef]
18. Walker, D.; Xie, H.; Yan, K.K.; Maslov, S. Ranking scientific publications using a model of network traffic. *J. Stat. Mech. Theory Exp.* **2007**, *2007*, P06010. [CrossRef]
19. Yao, L.; Wei, T.; Zeng, A.; Fan, Y.; Di, Z. Ranking scientific publications: The effect of nonlinearity. *Sci. Rep.* **2014**, *4*, 6663. [CrossRef]
20. Zhou, J.; Zeng, A.; Fan, Y.; Di, Z. Ranking scientific publications with similarity-preferential mechanism. *Scientometrics* **2016**, *106*, 805–816. [CrossRef]
21. Brin, S.; Page, L. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.* **1998**, *30*, 107–117. [CrossRef]
22. Su, C.; Pan, Y.; Zhen, Y.; Ma, Z.; Yuan, J.; Guo, H.; Yu, Z.; Ma, C.; Wu, Y. PrestigeRank: A new evaluation method for papers and journals. *J. Inf.* **2011**, *5*, 1–13. [CrossRef]
23. Mariani, M.S.; Medo, M.; Zhang, Y.C. Identification of milestone papers through time-balanced network centrality. *J. Inf.* **2016**, *10*, 1207–1223. [CrossRef]
24. Shen, H.W.; Barabási, A.L. Collective credit allocation in science. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 12325–12330. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.