Can we use link-based indicators to find highly cited publications? The case of the Trust Flow score

Enrique Orduña-Malea[1]
Isidro F. Aguillo[2]

[1] Department of Audiovisual Communication, Documentation and History of Art, Universitat Politècnica de València, Valencia, Spain
✉ enorma@upv.es
[2] Cybermetrics Lab, Institute of Public Goods and Policies (IPP), Spanish National Research Council (CSIC), Madrid, Spain
✉ isidro.aguillo@csic.es

Abstract
The Majestic's Trust Flow (TF) is a link-based score aimed at measuring the influence of online objects (e.g., scientific publications) by considering the weighted number of links received from trusted websites. This study describes the bibliographic characteristics and impact of those publications with the highest TF score. In order to do this, 20,810 URL-based DOIs were identified and analysed. The results show that these DOIs mainly represent recent publications (57.1% of publications were published between 2010 and 2020), journal articles (93.75%) published in the first SJR quartile (81.7%), written with international collaboration (40.4%) and biased towards the field of Medicine (36.9%). While the TF score is a discovering tool with the potential to be used in webometric studies to find influential publications, a few technical limitations jeopardize the general applicability of this indicator for research evaluation at the publication level.

Keywords
Link analysis; research impact; bibliometrics; webometrics; influential publications; altmetrics

This manuscript is an accepted copy version. Please cite as follows:

Orduña-Malea, E., & Aguillo, I. F. (2025). Can we use link-based indicators to find highly cited publications? The case of the Trust Flow score. *Journal of Information Science*. https://doi.org/10.1177/01655515221141032

1. Introduction

Highly cited publications (including all document types) are documents generally considered of high significance, utility and novelty, and being presumably much more influential than other publications [1]. These publications hold specific bibliographic properties [2], such as being co-authored by many researchers, often involving international collaboration, receiving citations from a large number of different publication sources and exhibiting generally higher interdisciplinarity [3]. These specific features make highly cited publications important units of analysis in bibliometrics, not only to find research excellence at the aggregate [4] or individual level [5], but also to identify the main influential documents to the scientific progress [6].

However, there is no consensus on the exact definition of highly cited publications [7, 8], and different methods to determine the status of "highly cited" can be used. For example, we can use absolute thresholds (i.e., a minimum number of citations received), relative thresholds (e.g., the Essential Science Indicators database uses citation percentiles to identify highly cited journal articles according to the year of publication and the discipline), distance statistics (e.g., the number of citations received is at least k times the mean citation rate of the publication source in which the publication was published [9]) or a combination of requirements. For instance, Google Scholar's Classic Papers required publications to be journal articles or conference communications published in one specific year, among the top 10 most cited documents in their respective subject category, attaining a minimum of 20 citations and written in English [10].

Nonetheless, measuring citation counts raises some undesired effects. The significance of the citations introduces theoretical concerns [11, 12], as extreme highly citation counts can be due to historical reasons [13] or might follow an imitation pattern [14]. Otherwise, other technical side effects are the over-representation of review papers, the dependence on the discipline, year of publication, document typology, and bibliographic database used [10, 15-16]. Despite the theoretical and technical limitations shown above, the generation of highly cited publication lists allows meta-researchers to identify, map and describe major scientific discoveries, considering their citation count [17]. Therefore, the article "Protein measurement with the Folin phenol reagent", by Lowry et al. [18], is still the most cited publication as of January 2022, a position that this journal article holds for several years [19], evidencing certain stability in the upper part of the ranking.

Other alternative indicators can also be used to generate "highly-X lists", with the general purpose of evidencing wider impacts of research publications. For example, when using Dimensions database, we can generate a list of the research publications achieving the highest Altmetric Attention Score (AAS) [20]. As of 6 January 2021, the top 10 publications with the highest AAS are all related to the COVID-19 pandemic, evidencing that this indicator is highly dynamic, reflecting current social and broad interest on scientific results and not necessarily related to the citation-based impact. In particular, the article "Covid-19: Researcher blows the whistle on data integrity issues in Pfizer's vaccine trial", by Thacker [21], is the publication with the highest AAS in Dimensions, but attaining few citations (16 citations in Google Scholar; three in Dimensions and Web of Science Core Collection; two in Scopus). Considering the date of publication (November 2021), the number of citations received might presumably increase.[1]

It goes without saying that altmetrics data have also come through several limitations, such as data volatility, reliability, quality and superficiality [22-25], the dependence on the altmetrics data source coverage [26, 27] and the effective use of the Web and social media platforms by researchers [28, 29]. Despite these limitations, altmetrics exhibit potential to fast delivering of indicators signalling wider impact of research [30], offering lists of highly disseminated publications covering controversial results or topics of public interest.

Furthermore, link counts (i.e., URL citations) can also be used to quantify the online impact of research publications. This metric is closely related to the altmetrics movement (e.g., tweet counts measure the number of tweets including a link to one publication). However, academic link analysis and altmetrics have studied link counts under different conceptual frameworks. While the altmetrics field focuses on those links from social media platforms to publications, the link analysis field focuses on links from websites to aggregated academic websites, such as universities [31, 32] or journals [33].

The application of the link analysis at an article-level has been limited due to historical technical limitations [34], especially the lack of dedicated link data sources, link spam and link inflation [35], and the later necessity of demanding cleansing data processes. Consequently, to the best of the authors' knowledge, there are no lists of the highly linked scientific publications.

In order to avoid the methodological weaknesses of link counts, the use of advanced indicators is advisable, such as counting referring domain names or links from trusted websites. In fact, search engine optimization (SEO) algorithms, such as the Google's PageRank [36] or HITS (Hyperlink-Induced Topic Search) [37], are based on rewarding those links received from prominent websites.

Unfortunately, general search engines with extensive link coverage (e.g., Google) do not provide link-based advanced indicators or data export features, jeopardizing the use of these databases for academic link analysis. Alternatively, commercial SEO tools (e.g., SEMRush, Ahref, Link Explorer or Majestic) offer a wide range of link-based indicators, although their link coverage is lower than Google's. Among these tools, Majestic stands out, covering a comprehensive database of URLs.

In 2012, Majestic introduced an advanced web indicator, Trust Flow (TF), which intends to measure the influence of domain names and webpages by considering the influence of those webpages that link to them (see Table 1 for further details). Other equivalent metrics exist in the web SEO market, such as Domain Authority (provided by Link Explorer) or Authority Score (provided by SEMrush). The rationale of the TF is that trusted websites tend to link to other trusted websites. Accordingly, websites closely linked to a trusted seed site can obtain higher scores, whereas websites having questionable links would obtain a much lower score. The TF has already been applied in webometric studies [38, 39], evidencing its usefulness to filter out dubious content and link spam when it comes to evaluate scholarly and scientific websites. However, no large-scale studies at the publication level have been carried out yet.

By testing the performance of the TF at the publication level, meta-researchers would broaden their understanding of the properties of highly cited publications. In addition, practitioners and other actors involved in research evaluation exercises would obtain useful information about the advantages and disadvantages of using this indicator to evaluate broader impact of research, supplementing altmetrics analyses.

Having this in mind, this study intends to find out which publications achieve significant values of online trust (i.e., high TF values) on the one hand, and to check whether online trust is related to other dimensions of impact (citation-based impact and altmetrics) on the other. These issues would allow us to understand which scientific literature is covered when links from trusted websites to publications are counted, and whether lists of the highly linked publications differ from the lists of highly cited publications.

Formally, the main research objectives of this study are drawn below:

RO1. Identify the publications with the highest TF score and define their core bibliographic characteristics.
RO2. Determine the correlation between the TF and the number of citations and altmetrics received at the publication level.

2. Method

2.1. Publication dataset

The doi.org domain name was used as a baseline to collect publications. A Digital Object Identifier (DOI) is an alphanumeric string assigned to digital objects which can be encoded in specific application contexts, like URLs and with the HTTPS protocol. The DOI system was announced in 1997 and it is managed by the International DOI Foundation (IDF), being approved as an ISO standard in 2010 (ISO 26324:2012).[2] It is aimed to provide a persistent identification together with current information about the object [40]. DOIs can be assigned by DOI Registration Agencies (e.g., Crossref or Datacite), having assigned over 275 million DOIs to date, mainly to identify scientific publications.[3] URL-based DOIs are hosted in doi.org as subdirectories (e.g., https://doi.org/10.1016/j.ipm.2008.10.001), thus allowing link analysis.

The 30,000 internal webpages of doi.org with the highest TF value were collected through the Majestic API (Full API plan) as of 26th December 2021.

For this study, the Majestic's internal online search feature was used, setting up the fresh index and the domain-level analysis. The text chain "doi.org" was inserted as a search term, and the database returned all internal webpages (i.e., each DOI-based link) broken down by the TF value. All data were subsequently downloaded in CSV file format.

The fresh index covers all links crawled during the last four months, minimizing the effects of deleted links, allowing to obtain a picture of the contemporary web. It contains all links discovered and crawled from September 2021 to December 2021. This index covers 1,109 billion URLs found, which are used to calculate automatically the TF of each DOI-based URL.

All the collected URL-based DOIs were subsequently cleansed. Thus, duplicate DOIs, broken DOIs (e.g., dx.doi.org/10), publisher-level DOIs (e.g., doi.org/10.1128/), DOIs used as examples (e.g., doi.org/xx.xxx/yyyy), as well as internal pages not linked to scientific publications (e.g., doi.org/ISSN) were all eliminated.

As http and https versions related to the same DOI are, technically speaking, different links, only the https version was considered. Finally, a TF value of 10 was arbitrarily used as a minimum threshold in order to avoid low trusted digital objects. This process yielded a final set of 20,810 URL-based DOIs, which are referred to as "trusted online publications". For each DOI-based URL, the TF value, the number of external links received and the number of linking domain names were collected.

2.2. Data collection

In order to accomplish the RO1, the Crossref API (publication source, publication type and year of publication) and the Scival research analytics tool (Scimago Journal Rank quartile, research disciplines, and collaboration) were used to collect bibliographic data for each of the 20,810 DOIs obtained previously. As regards the RO2, the PlumX API was used to capture citation-based metrics and altmetrics. All data was collected as of 7th January 2022. All metrics captured for each DOI are described in the Table 1.

Table 1. Article level metrics: source and scope.

Metric	Source	Scope
Trust Flow	Majestic	Proprietary score on a log-based scale between 0-100 achieved by a digital object (in this case, a publication represented by a URL-based DOI). It is based on the number of hyperlinks received from trusted seed websites manually selected by Majestic across 800 categories.
Link counts	Majestic	Total number of links received by each publication referred by the DOI.
Linking domain name counts	Majestic	Total number of domain names providing at least one link to each publication referred by the DOI.
Citation count	PlumX	Number of citations received by each publication referred by the DOI.*
Blog counts	PlumX	The number of blog posts written about the publication referred by the DOI, from a curated list.**
News counts	PlumX	The number of news articles written about the publication referred by the DOI.
Reader counts	PlumX	The number of people who have added the publication referred by the DOI to their Mendeley library.
Tweet counts	PlumX	The number of tweets and retweets that mention the DOI.

*Unique citations from Airiti ACI, Chinese Citation database, Crossref, PubMed Central, PubMed Central Europe, RePEc, SciELO, Scopus, and SSRN.
** https://plumanalytics.com/plumx-expands-mentions-category-by-tracking-blogs/

2.3. Data analysis

Due to the skewed distribution of link-based metrics, Spearman's rho correlations [41] were used to measure the strength of association between citation-based and online-based metrics (RO2). All statistical tests were carried out through XLSTAT 2021.1.1.[4]

To exemplify the impact achieved by DOI-based URLs, the linking domain names were manually categorized into domains: academic (universities, journals, publishers, bibliographic databases, research societies, research projects, etc.), informative (e.g., media, magazines, websites offering information), commercial (e.g., companies), entertainment (e.g., music, movies, videogames), governmental (e.g., public departments and services) and social (e.g., social media platforms, online fora, wikis). Connectivity between the linking domain names and the DOI-based URLs were exemplified by networks built with Gephi 0.9.1.[5]

3. Results

3.1. Bibliographic properties

The URL-based DOIs dataset covers scientific literature back from 1693 (a scanned article published in Philosophical Transactions) together with a huge concentration of publications in the last years (88.6% of all publications are published after 2000). The TF scores obtained are moderate, with higher values for recent publications. Most DOIs obtain TF values ranging from 10-20 (89.7%), whereas only four DOIs overpass the value of 40 (Table 2).

Table 2 URL-based DOIs by decade and TF score.

Decade	Trust Flow				TOTAL	%
	10 to 20	21 to 30	31 to 40	41 to 50		
1900-1910	13	0	0	0	**13**	0.1
1910-1920	9	0	0	0	**9**	0.0
1920-1930	24	0	0	0	**24**	0.1
1930-1940	38	6	0	0	**44**	0.2
1940-1950	38	4	0	0	**42**	0.2
1950-1960	111	7	0	0	**118**	0.5
1960-1970	162	13	0	0	**175**	0.8
1970-1980	314	20	1	0	**335**	1.6
1980-1990	495	31	0	0	**526**	2.4
1990-2000	1,097	79	1	0	**1,177**	5.5
2000-2010	4,727	323	0	0	**5,050**	23.4
2010-2020	10,966	1,352	10	0	**12,328**	57.1
2021	1,369	373	9	4	**1,755**	8.1
TOTAL	**19,362**	**2,208**	**21**	**4**	**21,595**	**100**

Source: Crossref and Majestic.

URL-based DOIs basically refer to journal articles (93.75%), with the remaining document types being residual (Table 3). Among the journals publishing the most articles, we can find prestigious multidisciplinary journals (e.g., Nature, Science or PNAS), open access mega-journals (e.g., PLoS ONE, Nature Communications, Scientific Reports) and medical journals (e.g., Lancet, New England Journal of Medicine, JAMA, Cell or BMJ). The presence of journals specialized in statistics (e.g., Journal of Statistical Software, and Journal of the American Statistical Association) also stands out (Table 4).

Table 3 Distribution of trusted online publications according to publication type.

Document type	Number of trusted online publications	%
Journal article	19,257	93.75
Book chapter	313	1.52
Proceeding article	300	1.46
Posted content	297	1.45
Book	241	1.17
Monograph	55	0.27
Report	34	0.17
Other	12	0.06
Reference entry	9	0.04
Dataset	7	0.03
Journal	7	0.03
Reference book	3	0.01
Report series	3	0.01
Component	1	0.00
Proceedings	1	0.00
Standard	1	0.00
TOTAL	**20,541**	**100**

Source: Crossref

Table 4 Journals with most trusted online publications published.

Publication source	Number of trusted online publications	%
Nature	1,137	5.7
Science	704	3.5
Proceedings of the National Academy of Sciences	550	2.8
The Lancet	447	2.2
New England Journal of Medicine	441	2.2
PLOS ONE	275	1.4
JAMA	247	1.2
Journal of Statistical Software	222	1.1
Cell	218	1.1
BMJ	209	1.0
Nature Communications	155	0.8
Bioinformatics	139	0.7
Scientific Reports	110	0.6
Journal of the American Statistical Association	100	0.5
Nucleic Acids Research	100	0.5
The European Physical Journal C	100	0.5

Source: Crossref

Scival contains 80.6% (16,778) of all the publications referred to by a URL-based DOI with at least a TF value of 10. These publications are mainly published in journals ranked in the first quartile (81.7%), according to the Scimago Journal Rank (Figure 1), which evidence the capability of the TF metric to identify publications published in quality indexed journals.

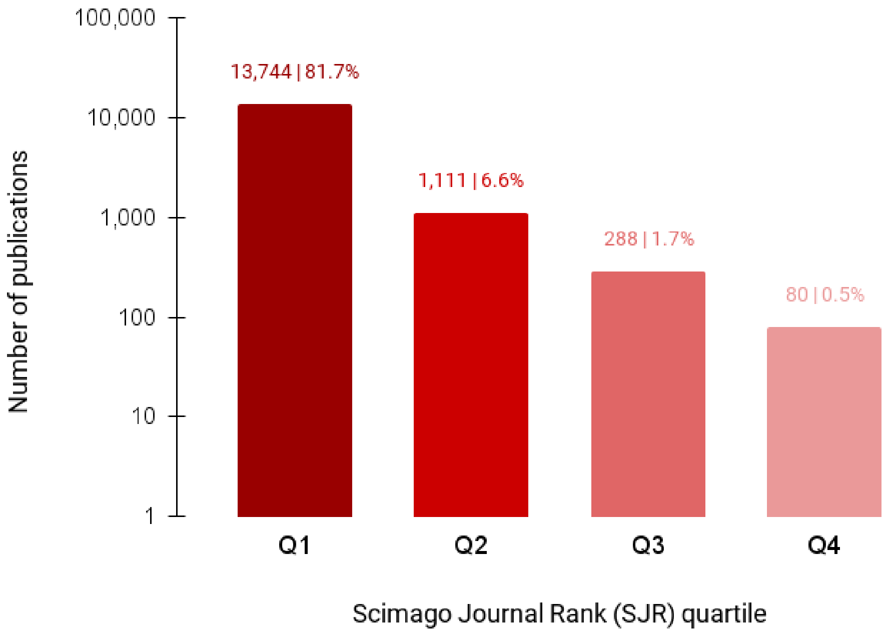


Figure 1 Trusted online publications per quartile (Scimago Journal Rank-SJR).
Source: Scival

As regards the co-authorship patterns, the results indicate that most trusted online publications (40.4%) are published with international collaboration, with publications authored by just one author being scarce (10.1%). The number of publications co-authored with private companies (7.1%) is significant, given the low global percentage of this type of collaboration (Figure 2).

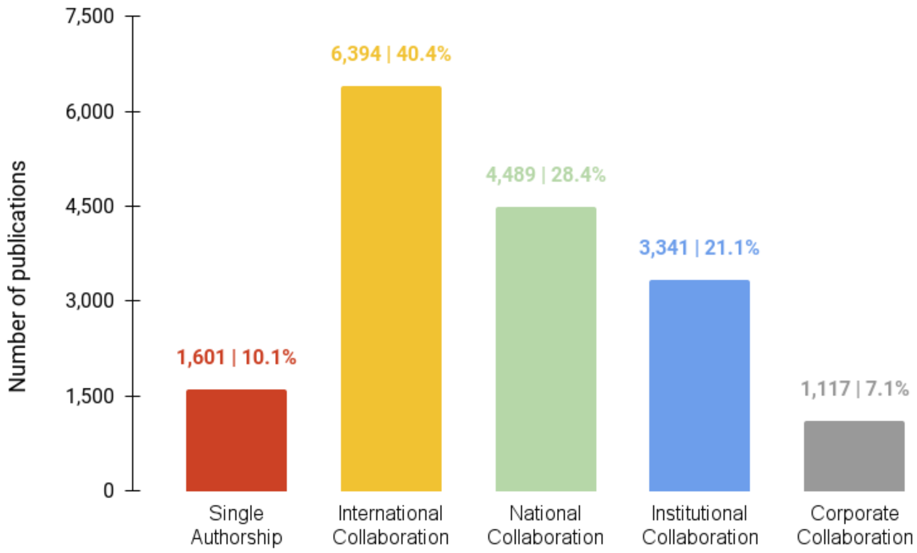


Figure 2 Collaboration patterns of the trusted online publications dataset (1996 to 2020).
Source: Scival

The main topics covered by trusted online publications are medicine (36.9% of publications), biochemistry, genetics and molecular biology (17.0%), and mathematics (10.3%). The general multidisciplinary topic also achieves a significant percentage (14.3%) of publications (Figure 3). The specific publication context during the COVID-19 pandemic is also reflected in the increase of trusted online publications in medicine (1,141 publications in 2020).

Comparing these results with the overall distribution in Scopus in that period (See supplementary material, Appendix A), we observe an overrepresentation of these disciplines in the dataset, while computer science, engineering, materials science, and physics and astronomy are significantly underrepresented.

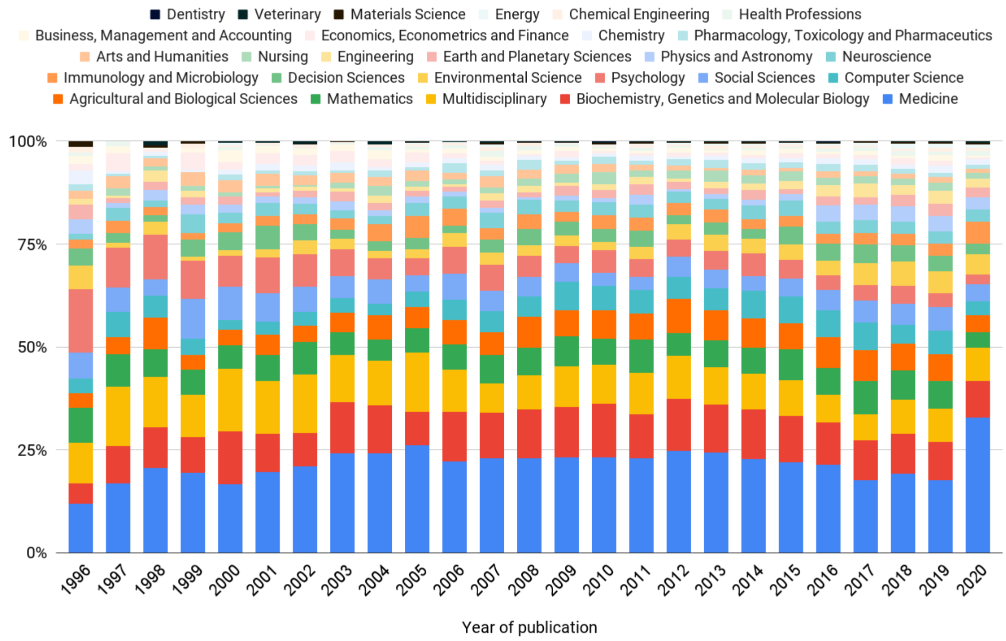


Figure 3 Disciplines covered by trusted online publications (1996 to 2020).
Source: Scival

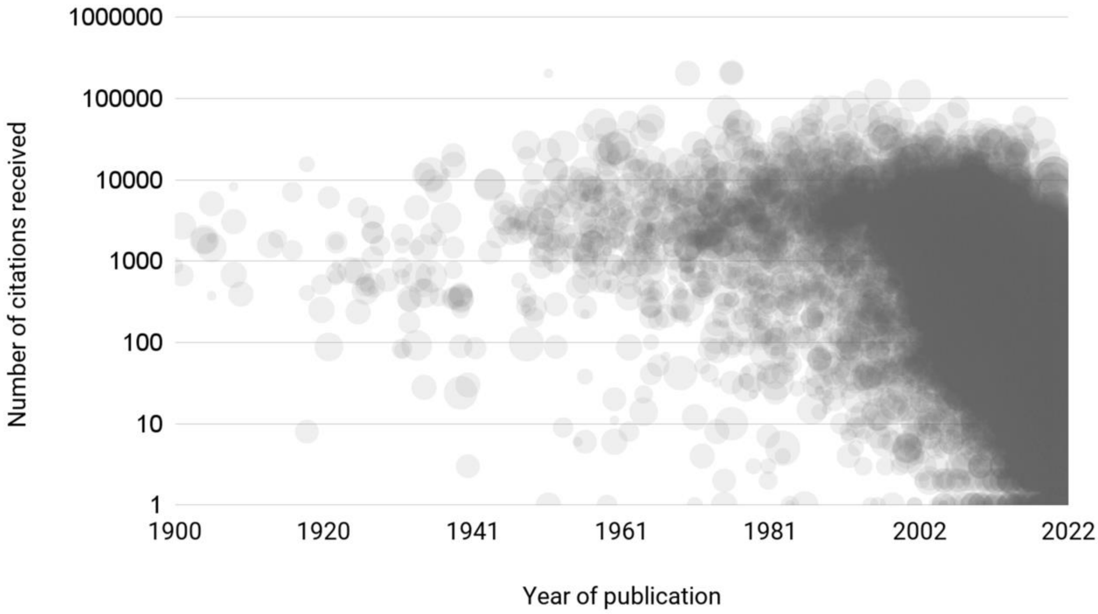
3.2. Impact of trusted online publications

Trusted online publications dataset contains a significant number of highly cited documents (26.9% of publications have attained at least 1,000 citations), achieving a median value of 227.5 citations (Table 5). Nevertheless, the dataset also contains publications with zero citations, most of them recently published with no time to attain citations yet (495 publications with zero citations were published in 2021). If we break down the publications according to citation ranges, the average and median TF scores oscillate between 15~18 values, being higher for those documents in the citation range of 50,000-to-100,000 citations received.

Table 5 Distribution of citation counts for trusted online publications.

Citation counts	Number of trusted online publications	%	Mean TF	Median TF
0	1,093	5.3	15.8	16.0
1 to 10	2,251	10.9	15.3	15.0
11 to 50	2,830	13.7	15.2	15.0
51 to 100	1,772	8.6	15.4	15.0
101 to 500	4,991	24.1	15.6	15.0
501 to 1000	2,184	10.6	15.7	15.0
1001 to 2000	2,140	10.3	15.4	15.0
2001 to 5000	2,184	10.6	15.4	15.0
5001 to 10000	803	3.9	15.9	15.0
10001 to 50000	407	2.0	17.1	17.0
50001 to 100000	17	0.1	18.2	18.0
> 100000	6	0.0	16.2	16.0
TOTAL	**20,678**	**100**		

Considering the period 2000-2020, the TF score captures publications regardless of the number of citations received (Figure 4a). Interestingly, most of the publications published in the 1950-2000 period receive an outstanding number of citations. TF values are generally higher for recent publications, regardless of the number of citations received, evidencing a recency effect (Figure 4b).



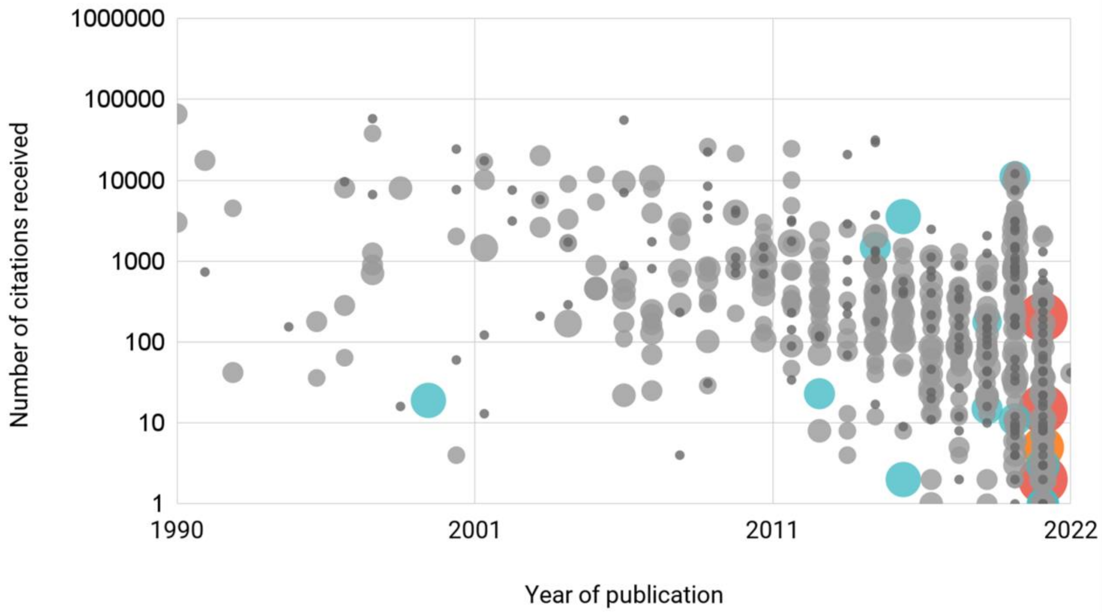


Figure 4 Scatterplot of the year of publication and number of citations received by trusted online publications dataset:
(a) Including all publications in the period; (b) Including publications with a TF value of at least 25 in the period.
Note: bubble size represents the TF value.

The recency effect determines the correlations obtained between the publication-level metrics. Table 6 includes different scenarios (all years, 5-year periods, last year) achieving diverse results per year. However, a general weak correlation between the TF metric and the remaining metrics is evidenced. Otherwise, the number of blogs/news/tweets is strongly correlated, especially in the last available year (2021). The number of citations received is strongly correlated to the number of Mendeley readers and the total number of external links.

Table 6 Spearman correlation values between publication-level metrics.

Data range= all years N= 6,544	A	B	C	D	E	F	G	H
Link counts (A)	1							
Domain name counts (B)	*0.58	1						
Trust Flow (C)	*0.26	*0.34	1					
Blog counts (D)	*0.18	*0.25	*0.23	1				
Citation counts (E)	*0.49	*0.16	-0.02	*-0.11	1			
News counts (F)	*0.05	*0.28	*0.28	*0.61	*-0.30	1		
Reader counts (G)	*0.51	*0.24	*0.07	0.00	*0.86	*-0.20	1	
Tweet counts (H)	*0.07	*0.17	*0.23	*0.52	*-0.25	*0.49	*-0.10	1

Data range= 2017 to 2021 N= 2,563	A	B	C	D	E	F	G	H
Link counts (A)	1							
Domain name counts (B)	*0.62	1						
Trust Flow (C)	*0.35	*0.37	1					
Blog counts (D)	*0.27	*0.26	*0.16	1				
Citation counts (E)	*0.44	*0.24	*0.22	*0.11	1			
News counts (F)	*0.22	*0.34	*0.21	*0.69	*0.05	1		
Reader counts (H)	*0.44	*0.27	*0.22	*0.13	*0.91	0.03	1	
Tweet counts (I)	*0.23	*0.20	*0.16	*0.46	*0.17	*0.43	*0.18	1

Data range= 2014 to 2018 N= 1,902	A	B	C	D	E	F	G	H
Link counts (A)	1							
Domain name counts (B)	*0.54	1						
Trust Flow (C)	*0.23	*0.29	1					
Blog counts (D)	*0.24	*0.21	*0.13	1				
Citation counts (E)	*0.53	*0.18	*0.09	*0.09	1			
News counts (F)	*0.15	*0.26	*0.21	*0.60	0.00	1		
Reader counts (H)	*0.52	*0.22	*0.12	*0.16	*0.89	0.03	1	
Tweet counts (I)	*0.22	*0.13	*0.07	*0.56	*0.20	*0.36	*0.26	1

Data range= 2021 N= 307	A	B	C	D	E	F	G	H
Link counts (A)	1							
Domain name counts (B)	*0.59	1						
Trust Flow (C)	*0.28	*0.33	1					
Blog counts (D)	*0.28	*0.22	*0.22	1				
Citation counts (E)	*0.29	*0.17	*0.32	*0.18	1			
News counts (F)	*0.31	*0.43	*0.33	*0.62	*0.26	1		
Reader counts (H)	*0.25	*0.15	*0.26	*0.26	*0.86	*0.24	1	
Tweet counts (I)	*0.20	*0.18	*0.21	*0.56	*0.26	*0.49	*0.31	1

* Values different from zero with a significance alpha-value = 0.05

The recency effect is evidenced when it comes to determining the trusted online publications with the highest TF value (See supplementary material, Appendix B). We observe that seven out of the top ten publications were published in 2021. The article which holds the first position corresponds to the article by Thacker [21], which also holds the highest AAS, as mentioned previously.

Beyond the COVID-19 related publications (with a huge number of tweet counts), we observe the presence of research publications covering topics of public interest related to climate change, history or psychology, but with limited attention on Twitter. Among these top trusted online publications, the most cited article is the "FAIR Guiding Principles for scientific data management and stewardship". However, if we focus the analysis on the most cited publications collected in the dataset (See supplementary material, Appendix C), we can identify those publications most cited in history, which also exhibit a remarkable number of Mendeley Reader counts, but a lack of current mentions from Twitter.

In order to test the capacity of the TF indicator to capture highly cited publications, the top 2,000 most cited publications according to Scopus have been used as a testbed, obtaining 1,879 publications with a DOI. Of these, 58.3% (1,091 publications) are included in the trusted online publications dataset. This percentage rises to 68.8% when only the highly cited publications published from 2000 onwards are considered; and to 79.5% for highly cited publications published from 2010 until now.

3.3. Origin of the linking trusted websites

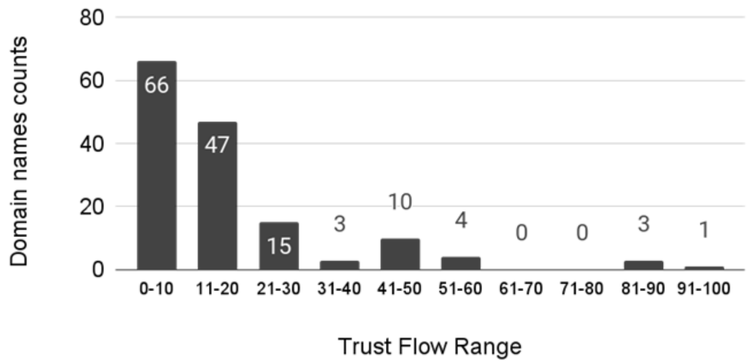
Two publications have been considered as a case study to better understand the impact provided by the TF score: the publication with the highest TF score (Figure 5) and the publication with most citations received (Figure 6). All the linking domain names to these DOI-based URLs have been categorized, considering link-data as of 14 April, 2022.[6]

The publication with the highest TF score receives 930 links from 149 linking domain names (Figure 5). 63.1% of the linking domain names come from informative websites, especially from blogs and media. Non-available websites (e.g., delorgeril.info, which provided 179 links) highlight the volatility of link data. Links from academic websites are scarce (211 links from 22 link domain names), a result that is aligned to the low number of citations received by this publication at the time of writing this paper.[1] Generally, the trustworthiness of the linking websites is low (75.8% of linking domain names show a TF value lower than 20). However, four linking domain names (those corresponding to Wikipedia, the journal BMJ, and the WordPress and BlogSpot blogs providers) are highly influential (they hold a TF score higher than 80).

On the contrary, the most cited publication shows an entirely different profile, receiving 3,659 links from 213 linking domain names (Figure 6). In this case, informative linking domain names are scarce while academic linking domain names are dominant (3,547 links from 191 linking domain names), especially from journals and publishers (3,300 links). The trustworthiness of the linking websites is different. In this case, 57.7% of the linking domain names hold a TF score lower than 20, but only one (nature.com) overpass the value of 80.

These results allow us to clearly observe the duality of the TF as an indicator. On the one hand, recent publications achieve higher TF scores due to the existence of links from informative websites (including media), which may be unavailable for a short time. Most links come from sites with low TF, but links from few highly (current) influential sites skyrocket the publication's TF score. On the other hand, older publications achieve moderate TF scores, due to the existence of links from academic publications, more stable over time, and which have a moderate-high TF score. The absence of links from highly (current) influential sites ultimately determines the moderate TF score obtained.

Domain names	Category	Link counts	Trust Flow
delorgeril.info	No available	179	42
alecomm.com	Informative - Personal website	157	4
altmetric.com	Academic - Database	76	51
bmj.com	Academic - Publisher	47	81
ralphdejongh.com	Entertainment	47	14
rxisk.org	Commercial	46	21
wakkeren.nl	Informative - Entity website	19	9
davidhealy.org	Academic - Personal website	19	16
wordpress.com	Unknown	15	84
nicoledelepine.fr	Academic - Personal website	15	37



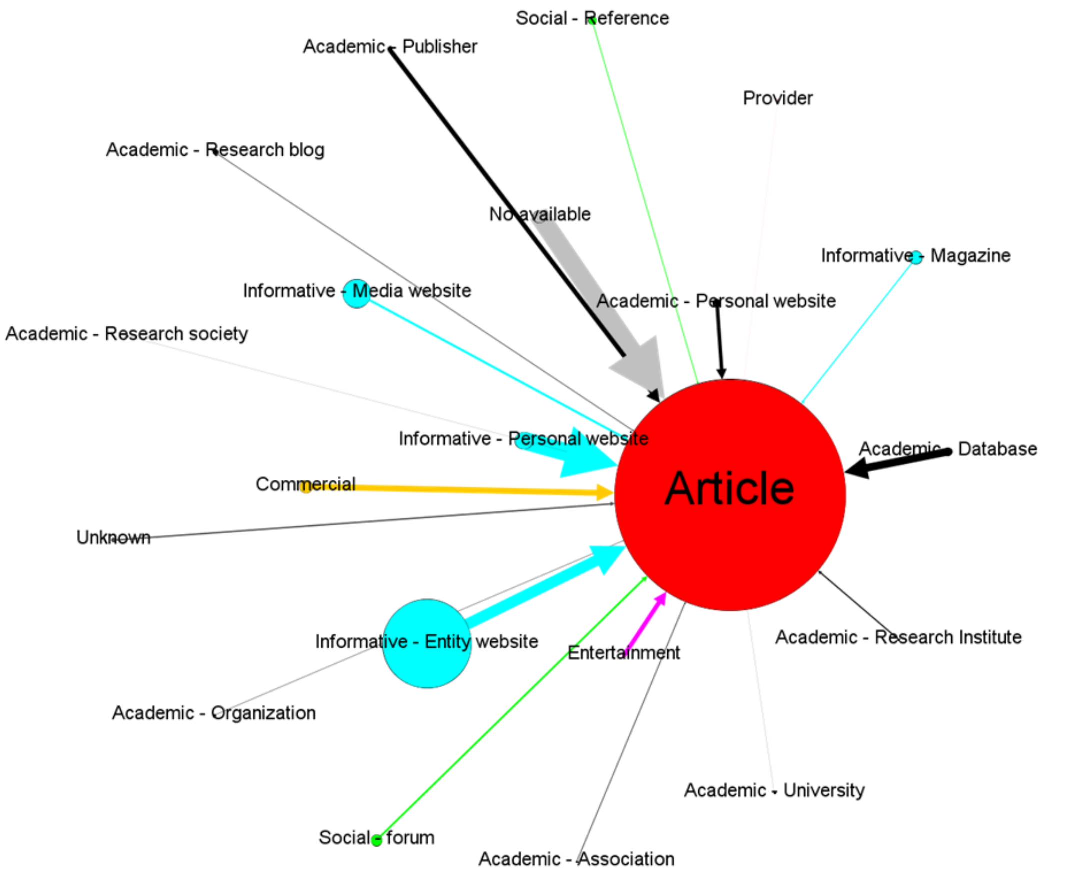
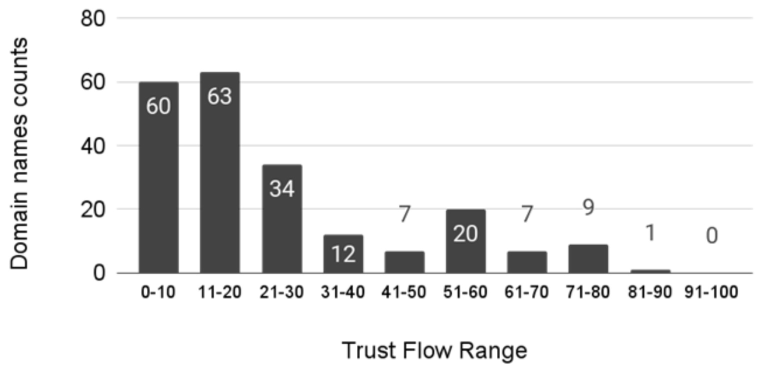


Figure 5 Publication sheet 1: the case of the article with the highest TF score (10.1136/bmj.n579).
Data as of 14th April 2022

Domain names	Category	Link counts	Trust Flow
biomedcentral.com	Academic - Publisher	1,110	74
springeropen.com	Academic - Publisher	250	34
frontiersin.org	Academic - Publisher	217	47
hindawi.com	Academic - Publisher	159	62
springer.com	Academic - Publisher	147	77
nature.com	Academic - Publisher	144	85
microbiologyresearch.org	Academic - Publisher	117	25
mdpi.com	Academic - Publisher	111	60
peerj.com	Academic - Publisher	69	30
aimspress.com	Academic - Publisher	69	26



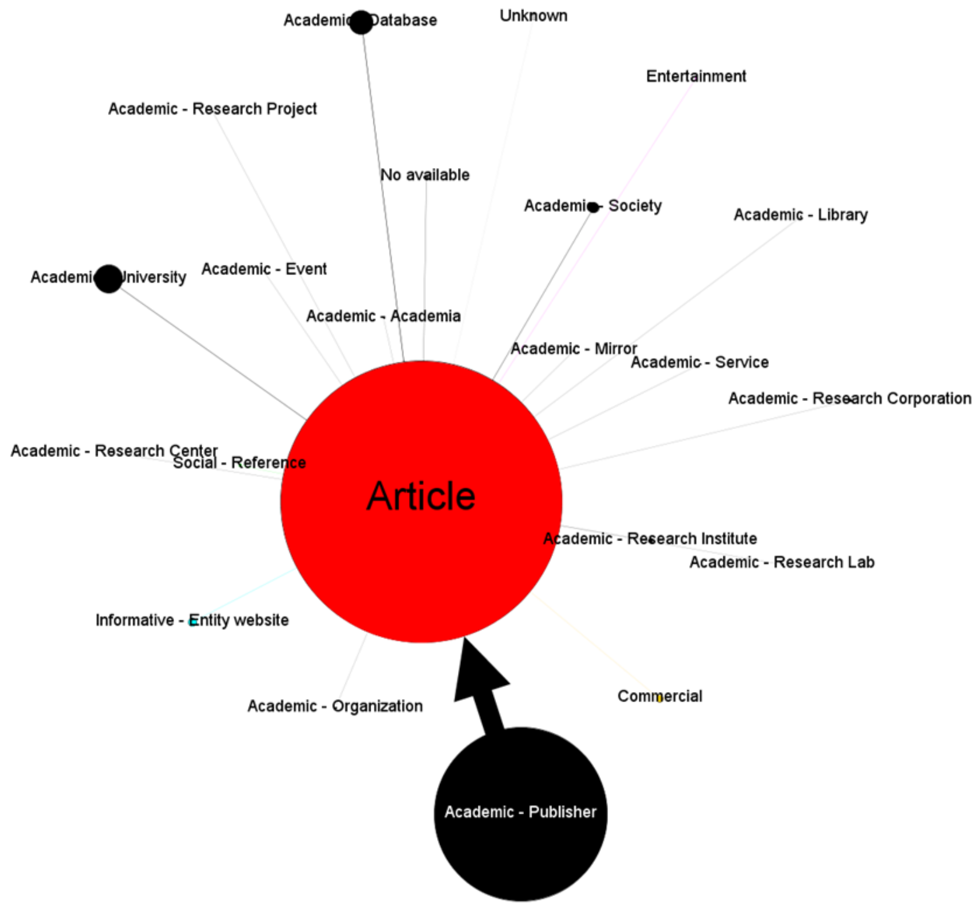


Figure 6 Publication sheet 2: the case of the article with most citations received (10.1016/0003-2697(76)90527-3)
Data as of 14th April 2022

4. Discussion

This work describes the scientific publications holding the highest TF scores, an online indicator that measures the influence of webpages and domain names according to the number of links received from external trusted websites, to determine the properties of the TF score for webometric, altmetric and bibliometric studies. This section discusses the results obtained.

4.1. Trusted online publications

As regards the core bibliographic characteristics of the trusted online publications (RO1), the TF score provides a limited set of publications (20,810) with at least a TF score of 10. These publications consist mainly of journal articles published in prestige journals (mainly indexed in the first quartile according to the Scimago Journal Rank), written with international or national collaboration, and with a broad bias towards the fields of medicine, biochemistry and statistics. The representation of disciplines offered by the TF differs from that offered by the Scopus database, evidencing the singularity of the indicator.

The impact of the trusted online publications can be labelled into two different dimensions (RO2):

Emergent publications: this set is composed of recent publications with an outstanding online visibility but a great variability on the number of citations received. These publications are candidates to become highly cited publications. For these publications, news counts, blog counts and tweet mentions do moderately correlate among them.

Classic publications: this set is composed of old publications, with many citations received, among which we can find the most cited articles in history (according to Scopus), and that shape the backbone of scientific discoveries of the previous century. For these publications, the number of citations is strongly correlated to the number of Mendeley's reader counts, evidencing that users currently include these classics in their reference managers (i.e., they are still citing the publications). Social media visibility of these publications is low.

4.2. Implications of the study

Science studies researchers can use TF as an advanced link-based indicator capable of measuring the online impact of publications, overcoming well-known limitations of raw links counts [35]. As this indicator relies on the existence of a URL associated with the publication, its use is independent of the language, publication type or discipline. Scientometricians can use the TF values associated to publications to better understand the current impact of old publications, to discover the emergence of new influential works, to monitor the online impact of publications over time, or to discover and describe the web sphere (both academic and non-academic) that links to the publications. Therefore, the opportunity is opened for many studies to better understand the dissemination and consumption of research publications. In addition, these analyses can be applied to other URLs beyond the DOI-based link, and even study other aggregations, especially repositories and journals.

On the other hand, the indicator can be used as a publication discovery tool by practitioners. By relying on links from reputable sites, the TF provides information not incorporated in alternative metrics (altmetrics), complementing them. Therefore, the TF could be cautiously used to obtain alternative evidence of impact and elaborate professional metric reports that incorporate the web dimension beyond social networks and web usage metrics.

4.3. Limitations of the study

The research baseline (DOIs): this study is limited to publications with an assigned DOI. This criterion generates a bias towards recent publications (old publications might not have DOIs, while almost recent publications do). In addition, previous studies have highlighted a few limitations of the DOIs. For example, URL-based DOIs can disappear [42], be wrongly assigned to publications [43-45] or show typographic errors [46, 47]. This study has identified similar limitations. For example, the following DOIs (10.1016/0003-2697(76)90527-3 and 10.1006/abio.1976.9999) refer to the same publication [48], which is precisely the most cited article in the dataset and the second most cited publication according to Scopus.

The analytical framework (*Majestic)*: The TF score of each URL-based DOI has been used as a method to collect trusted online publications. The coverage of Majestic is critical to ensure that data from as many publications as possible are processed. In this sense, Majestic surveys the Internet by offering one of the largest commercial link intelligence databases. Majestic collects 5,392,460,611 external links from 852,801 linking domain names related to the "doi.org" domain name as of 2nd January 2022, which is considered a large enough coverage for the purposes of this work. Otherwise, the Majestic's Fresh Index has been used. While this decision biases the results towards recent publications, on the one hand, it offers a more credible view of the current influence and impact of the websites, on the other. Future studies should analyse Majestic's historic index to check whether the biases found towards recent publications are minimized.

The use of the TF at the URL-level comes with some limitations that should be acknowledged. First, http and https versions are treated as distinct URLs, which diminish the real TF value that both the webpage (in this case, the publication) and the domain name (linking website) could achieve. Second, publications might have alternative URLs and IDs (e.g., website URLs, repository handles) which also scatter the reception of external links, reducing consequently the final TF score achieved. Third, the TF is biased towards commercial domain names (e.g., big companies and media) which receive a great number of shallow links (e.g., links to websites' homepages mainly due to recognition and promotional motivations), obtaining high TF scores. Conversely, links to URL-based DOIs are deep links (i.e., links to specific webpages mainly motivated by their contents), which are less bulky. Fourth, the TF is a composite indicator whose full methodology is unknown (e.g., the list of trusted seed sites is not available), jeopardizing its transparency and reproducibility. The only variable used to calculate the TF is the number of links received by each webpage, while the weights used to each link and the final transformation to a log-based scale remain unknown. Fifth, the average TF scores differ from the industry ones. As a publication's TF score depends on the TF of the domain names linking to the publication, being visible in certain industries (e.g., media) might determine the TF score obtained.

Bibliographic data (Crossref): the API provided publication data for 98.5% of all DOIs analysed, covering practically all the publications under study. These 262 DOIs with no information were mainly related to artefacts that did not correspond to publications themselves (e.g., figures, supplementary material), DOIs not registered by Crossref or URL-based DOIs online unavailable, which were filtered out. Crossref did provide a limited number of records with missing or wrong information in the publication date and publication source. These records were manually cleansed when possible.

Bibliographic data (Scival): While research has found that 1996 is no longer a watershed year for Scopus coverage [49], Scival data is still restricted to publications from 1996 onwards. The analysis carried out covered the 84.2% of all DOIs analysed (17,529), evidencing that most

DOIs collected by the TF correspond to recent publications indexed in Scopus. These results are considered representative enough of the whole dataset.

Altmetrics data (PlumX): The API provided data for the 99.37% of the DOIs analysed (20,680). The few DOIs with no response from the API were mainly due to the existence of URL-based DOIs online unavailable. Despite the availability of other altmetrics data providers, PlumX was used as it captures citation data from different bibliographic sources, including Scopus. Literature has shown that PlumX stands out providing data on Mendeley readers, while Altmetric.com provides more blog posts, news and tweet counts [26]. These differences between data providers should be considered when contextualizing the results obtained. Otherwise, some inconsistencies in citation data have been found. For example, PlumX API provided 203,833 citations to the historic article "Protein measurement with the Folin phenol reagent", by Lowry et al. [18], while Scopus provided 260,490. The computation of citations to this work has previously presented problems in other databases, such as Google Scholar [15].

Otherwise, the correlations between the TF scores and the citation and altmetrics data should be taken cautiously. The low prevalence of the altmetrics data [50] and the effect of time on correlations [51] have been already evidenced in the literature. In order to avoid the effect of null and zero values (e.g., 25.1% of the DOIs returned no data on tweet counts, and the 24.4% returned 0 mentions), all the Spearman correlations were calculated with publications holding at least a value of 1 in all the variables. This reduced the sample size, but it offered a less biased view of the publications' online behaviour.

As the TF score is logarithmically scaled from 0 to 100 and it relies mainly on current external links from still available trusted sites, this metric is not sensitive to the cumulative nature of the scientific publication. If one publication stops receiving new external links, its TF value will decrease, while recent publications will skyrocket their TF scores if they are highly linked (from trusted sites) in a short period of time.

This may explain why the number of citations received by the publications in the dataset achieved moderate correlations with the total number of links received, while the TF value achieved a weak correlation. As the links to DOIs mainly come from other publications (DOIs are included as part of the bibliographic reference), counting links can be of much value to publication-level analyses, whereas the TF and the linking domain names counts acquire more value to analyse aggregated websites (e.g., journals, universities), which receive more institutional and promotional links.

Finally, the concept of "trust" should be interpreted solely as the procedure used by Majestic to operationalize "website trust" through the Trust Flow indicator, which includes an element of public opinion engineering. Other link intelligence tools might use other methods. It is outside the purposes of this work to define what trust is or should be.

5. Conclusions

This study has carried out a large-scale analysis of URL-based DOIs to obtain evidence of the capacity of the TF score to identify influential publications.

The TF has been proved to be a discovering tool with potential to be used in webometric studies aimed at finding influential publications. However, this study has shown there is need to avoid or minimize certain limitations (e.g., multiplicity of publication URLs, data variability and obsolescence, reliance on private data providers, dependence on the referring domains industry), which jeopardize the general applicability of the TF for publication-level research evaluation. All these aspects should be delved into in future webometric studies.

The use of the TF (and equivalent indicators) to analyse the impact of aggregated units (e.g., journal websites, repositories and universities), which attain a wide variety of hyperlinks from trusted websites, also constitutes a promising line for academic link analysis.

Competing interests

The authors have no competing interests.

Funding information

This research has been funded by the Valencian Regional Government (Spain), through the research project UNIVERSEO (Ref. GV/2021/141).

Supplementary data

http://dx.doi.org/10.13140/RG.2.2.30421.33767

Notes

1. As of 16th April, the article receives 26 citations from Google Scholar, 7 from Dimensions, 5 citations from Web of Science Core Collection, and 4 from Scopus.
2. https://www.iso.org/standard/43506.html
3. https://www.doi.org/factsheets/DOIKeyFacts.html
4. https://www.xlstat.com
5. https://gephi.org/
6. While the original data were collected as of January, the two case studies were analysed in April 2022.

References

[1] Thelwall M. The influence of highly cited papers on field normalised indicators. *Scientometrics* 2019; 118: 519–537.
[2] Aksnes DW. Characteristics of highly cited papers. *Research Evaluation* 2003; 12: 159–170.
[3] Chen S, Qiu J, Arsenault C and Larivière V. (2021). Exploring the interdisciplinarity patterns of highly cited papers. *Journal of Informetrics* 2011; 15: 101124.
[4] Tijssen RJ, Visser MS and Van Leeuwen, TN. Benchmarking international scientific excellence: are highly cited research papers an appropriate frame of reference? *Scientometrics* 2002; 54: 381–397.
[5] Bornmann L and Leydesdorff L. (2018). Count highly-cited papers instead of papers with h citations: use normalized citation counts and compare "like with like"! *Scientometrics* 2018; 115: 1119–1123.
[6] Garfield E. Introducing Citation Classics: The human side of scientific papers. *Current Contents* 1977; 1: 5–7, http://www.garfield.library.upenn.edu/essays/v3p001y1977-78.pdf (accessed 17 April 2022).
[7] Bornmann L. (2014). How are excellent (highly cited) papers defined in bibliometrics? A quantitative analysis of the literature. *Research Evaluation* 2014; 23: 166–173.
[8] Glänzel W and Schubert A. (1992). Some facts and figures on highly cited papers in the sciences, 1981–1985. *Scientometrics* 1992; 25: 373–380.
[9] Glänzel W and Czerwon HJ. What are highly cited publications? A method applied to German scientific papers, 1980–1989. *Research Evaluation* 1992; 2: 135–141.
[10] Martín-Martín A, Orduna-Malea E and Delgado López-Cózar E. Coverage of highly-cited documents in Google Scholar, Web of Science, and Scopus: a multidisciplinary comparison. *Scientometrics* 2018; 116: 2175–2188.
[11] Cronin B. *The Citation Process: The Role and Significance of Citations in Scientific Communication*. Oxford: Taylor Graham, 1984.
[12] Moed HF. *Citation analysis in research evaluation*. Dordrecht (Netherlands): Springer Science & Business Media, 2006.

[13] Oppenheim C and Renn SP. Highly cited old papers and the reasons why they continue to be cited. *JASIS* 1978; 29: 225–231.
[14] Merton RK. The Matthew effect in science: The reward and communication systems of science are considered. *Science* 1968; 159: 56–63.
[15] Martín-Martín A, Orduna-Malea E, Ayllón JM and Delgado López-Cózar, E. A two-sided academic landscape: snapshot of highly-cited documents in Google Scholar (1950-2013). *Revista Española de Documentación Científica* 2016; 39: e149.
[16] Orduna-Malea E, Martín-Martín A and Delgado López-Cózar, E. Classic papers: using Google Scholar to detect the highly-cited documents. In: *Proceedings of the 23rd International Conference on Science and Technology Indicators—STI 2018*, Leiden, Netherlands, pp. 1298–1307. Leiden: Centre for Science and Technology Studies (CWTS), http://hdl.handle.net/1887/65275 (accessed 17 April 2022).
[17] Van Noorden R, Maher B and Nuzzo R. The top hundred papers. *Nature* 2014; 514, 550–553.
[18] Lowry OH, Rosebrough NJ Farr AL and Randall RJ. Protein measurement with the Folin phenol reagent. *Journal of biological chemistry* 1951; 193: 265–275.
[19] Kresge N, Simoni RD and Hill RL. The most highly cited paper in publishing history: Protein determination by Oliver H. Lowry. *Journal of Biological Chemistry* 2005; 280: e25, http://www.jbc.org/content/280/28/e25 (accessed 17 April 2022).
[20] Orduña-Malea E and Delgado-López-Cózar E. Dimensions: Re-discovering the ecosystem of scientific information. *El profesional de la información* 2018; 27: 420–431.
[21] Thacker PD. Covid-19: Researcher blows the whistle on data integrity issues in Pfizer's vaccine trial. *BMJ* 2021; 375: 1–3.
[22] Bornmann L. Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics* 2014; 8: 895–903.
[23] Sugimoto CR, Work S, Larivière V and Haustein S. Scholarly use of social media and altmetrics: A review of the literature. *JASIST* 2017; 68: 2037–2062.
[24] Thelwall M. (2020). The pros and cons of the use of altmetrics in research assessment. *Scholarly Assessment Reports*, 2020; 2: 2, https://wlv.openrepository.com/handle/2436/623132 (accessed 17 April 2022).
[25] Wouters P, Zahedi, Z and Costas R. (2019). Social media metrics for new research evaluation. In: Glänzel W, Moed HF, Schmoch U and Thelwall M. (eds.) *Springer handbook of science and technology indicators*, Cham: Springer, 2019, pp. 687–713.
[26] Ortega, JL. Reliability and accuracy of altmetric providers: a comparison among Altmetric. com, PlumX and Crossref Event Data. *Scientometrics*, 2018; 116: 2123–2138.
[27] Zahedi Z and Costas R. General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PloS ONE* 2018; 13: e0197326.
[28] Más-Bleda A and Aguillo IF. Can a personal website be useful as an information source to assess individual scientists? The case of European highly cited researchers. *Scientometrics* 2013; 96: 51–67.
[29] Más-Bleda A, Thelwall M, Kousha K and Aguillo IF. Do highly cited researchers successfully use the social web? *Scientometrics* 2014; 101: 337–356.
[30] Thelwall M and Kousha K. Web indicators for research evaluation. Part 2: Social media metrics. *Profesional de la Información* 2015; 24: 607–620.
[31] Ortega JL and Aguillo IF. Mapping world-class universities on the web. *Information Processing & Management* 2009; 45: 272–279.
[32] Thelwall M and Zuccala A. A university-centred European Union link analysis. *Scientometrics* 2008; 75: 407–420.
[33] Thelwall M. Journal impact evaluation: A webometric perspective. *Scientometrics* 2012; 92: 429–441.
[34] Thelwall M. (2021). The story of a colony: Theory development in webometric research. In: Sonnenwald DH. (ed.). *Theory development in the information sciences*. Texas: University of Texas Press, 2021, pp. 164–182.
[35] Thelwall M and Kousha K. Web indicators for research evaluation. Part 1: Citations and links to academic articles from the Web. *Profesional de la Información* 2015; 24: 587–606.
[36] Brin S and Page L. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems* 1998; 30: 107–117.

[37] Kleinberg JM. Authoritative Sources in a Hyperlinked Environment. In: Newman M, Barabási A.-L and Watts DJ. (Eds.). *The Structure and Dynamics of Networks*. Princeton: Princeton University Press, 2006, pp 604–632.
[38] Orduña-Malea E. (2021). Dot-science top level domain: Academic websites or dumpsites? *Scientometrics* 2021; 126: 3565–3591.
[39] Orduña-Malea E and Costas, R. Link-based approach to study scientific software usage: The case of VOSviewer. *Scientometrics* 2021; 126: 8153–8186.
[40] Paskin N. Digital object identifiers for scientific data. *Data science journal* 2005; 4: 12–20.
[41] Spearman C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* 1904; 15: 72–101.
[42] Ducut E, Liu F and Fontelo P. An update on Uniform Resource Locator (URL) decay in MEDLINE abstracts and measures for its mitigation. *BMC Medical Informatics and Decision Making* 2008; 8: 23.
[43] Franceschini F, Maisano D and Mastrogiacomo L. Errors in DOI indexing by bibliometric databases. *Scientometrics* 2015; 102: 2181–2186.
[44] Valderrama-Zurián JC, Aguilar-Moya R, Melero-Fuentes D and Aleixandre-Benavent R. A systematic analysis of duplicate records in Scopus. *Journal of Informetrics* 2015; 9: 570–576.
[45] Xu S, Hao L, An X, Zhai D and Pang H. Types of DOI errors of cited references in Web of Science with a cleaning method. *Scientometrics* 2019; 120: 1427–1437.
[46] Huang M and Liu W. Substantial numbers of easily identifiable illegal DOIs still exist in Scopus. *Journal of Informetrics* 2019; 13: 901–903.
[47] Zhu J, Hu G and Liu W. DOI errors and possible solutions for Web of Science. *Scientometrics* 2019; 118: 709–718.
[48] Bradford MM. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical biochemistry*, 1976; 72: 248–254.
[49] Thelwall M and Sud, P. Scopus 1900–2020: Growth in articles, abstracts, countries, fields, and journals. *Quantitative Science Studies* 2022; 3: 37–50.
[50] Costas R, Zahedi Z and Wouters P. Do "altmetrics" correlate with citations? Extensive comparison of altmetric indicators with citations from a multidisciplinary perspective. *JASIST* 2015; 66: 2003–2019.
[51] Thelwall M, Haustein S, Larivière V and Sugimoto CR. Do altmetrics work? Twitter and ten other social web services. *PloS ONE*, 2013; 8: e64841.