Sneaked references: Fabricated reference metadata distort citation counts

Lonni Besançon, Guillaume Cabanac, Cyril Labbé, Alexander Magazinov

Received: 3 October 2023 | Revised: 3 April 2024 | Accepted: 4 April 2024

DOI: 10.1002/asi.24896

RESEARCH ARTICLE

JASIST WILEY

Sneaked references: Fabricated reference metadata distort citation counts

Lonni Besançon[1] | **Guillaume Cabanac**[2,3] | **Cyril Labbé**[4] | **Alexander Magazinov**[5]

[1]Media and Information Technology, Linköping University, Norrköping, Sweden

[2]Université Toulouse 3 – Paul Sabatier, IRIT UMR 5505 CNRS, Toulouse, France

[3]Institut Universitaire de France (IUF), Paris, France

[4]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, Grenoble, France

[5]Yandex.Kazakhstan, Almaty, Kazakhstan

Correspondence
Guillaume Cabanac, Université Toulouse 3 – Paul Sabatier, IRIT UMR 5505 CNRS, 31062 Toulouse, France.
Email: guillaume.cabanac@univ-tlse3.fr

Funding information
European Research Council, Grant/Award Number: 951393

Abstract

We report evidence of an undocumented method to manipulate citation counts involving "sneaked" references. Sneaked references are registered as metadata for published scientific articles in which they do not appear. This manipulation exploits trusted relationships between various actors: publishers, the Crossref metadata registration agency, digital libraries, and bibliometric platforms. By collecting metadata from various sources, we show that extra undue references are actually sneaked in at Digital Object Identifier (DOI) registration time, resulting in artificially inflated citation counts. As a case study, focusing on three journals from a given publisher, we identified at least 9% sneaked references (5978/65,836) mainly benefiting two authors. Despite not being present in the published articles, these sneaked references exist in metadata registries and inappropriately propagate to bibliometric dashboards. Furthermore, we discovered "lost" references: the studied bibliometric platform failed to index at least 56% (36,939/65,836) of the references present in the HTML version of the publications. This research led to an investigation by Crossref (confirming our findings) and to subsequent corrective actions. The extent of the distortion—due to sneaked and lost references—in the global literature remains unknown and requires further investigations. Bibliometric platforms producing citation counts should identify, quantify, and correct these flaws to provide accurate data to their patrons and prevent further citation gaming.

1 | INTRODUCTION

It is now well recognized that the *Publish or Perish* atmosphere fuels questionable research practices (Crous, 2019). The introduction and widespread adoption of computed indicators (*h*-index, impact factor, etc.) have been leading academics to a situation where publishing is not enough and being cited is crucial. In this world of *Be Cited or Perish*, motivations for citation manipulations are on the rise (Lawrence, 2007). Possibilities of such manipulations have been documented by whistleblowers and researchers alike (Baccini et al., 2019; Haley, 2017).

This article is a revised version of a preprint posted on *arXiv* on October 3, 2023 (see https://doi.org/10.48550/arXiv.2310.02192).

Beel and Gipp (2010) experimented hiding citations from human eyes by using "white on white" text. Labbé (2010) achieved *h*-index manipulation through injection of meaningless texts containing a fixed set of references. Delgado López-Cózar et al. (2014) reproduced Labbé's experiment, demonstrating how the *h*-index and impact factors of real researchers and journals can be manipulated. It is worth noting that some editorial practices are similar to this type of manipulation: a seemingly legitimate editorial could cite all articles from a journal, thereby increasing its Impact Factor (e.g., Foley & Valkonen, 2012; Heathers & Grimes, 2022). Another method is the so-called "citation cartel" method (Franck, 1999). As part of the cartel, you cite specific authors who will cite you in return. This kind of manipulation also arises at the journal level (Davis, 2016; Kojaku et al., 2021). Another example is called "citation plantation"[1] and refers to undue over-citation of certain authors, even on unrelated topics. Last but not least, one of the most famous and common methods, is the addition of references during the peer-review process. Authors may be asked by reviewers and editors to add undue references to their submission. Whistleblowers and academic sleuths often try to detect citation manipulations through skews in citation (or self-citation) data (Szomszor et al., 2020; Van Noorden, 2020b; Wren & Georgescu, 2022).

As the motivation for and practice of citation manipulation gain traction, the consequences of such a practice are starting to emerge in academia. From time to time, highly cited researchers are banned from editorial boards (Van Noorden, 2020a) because they behave unethically by trading citations for manuscript acceptance. In 2021, Clarivate excluded 300 researchers from its *Highly Cited Researchers* list, and about 550 in 2022 (Oransky, 2022). This decision was taken based on evidence of citation manipulation. Another example: some malevolent individuals created hijacked journals by imitating current or defunct journals (Abalkina et al., 2022). They publish non-peer reviewed articles which cite other research works, leading to potential undue citations. Some manage to get these indexed by Elsevier's Scopus, a bibliometric platform which computes author-level indicators for research assessment (Baas et al., 2020).

It is worth pointing out that citation manipulation by various actors occurs in many places and at different times during the life cycle of a scientific publication. Until now, the documented manipulations always implied modifications of the version of record (Hinchliffe, 2022) (i.e., the published article available in PDF/HTML in its final version) by adding references to it. In this paper, we document a new loophole which is currently exploited: *sneaking undue references* during the DOI registration process by supplying additional and irrelevant fabricated metadata. The scientific publication itself, namely the version of record, remains unaltered and undue citations are actually *unreachable* by readers. We provide evidence that this manipulation is in use as we discovered in at least three journals of an open access publisher. Furthermore, our preprint version of this manuscript (Besançon et al., 2023) led to some media coverage (Singh Chawla, 2023) and triggered an independent investigation from Crossref which confirmed our findings. This not only highlights the potential of preprints to quickly advance science (Puebla et al., 2021) but also resulted in a quick correction from the responsible organization. This loophole will continue unless the metadata deposited by publishers are checked carefully.

2 | THE MANIPULATION: INCREASED CITATION COUNTS WITH SNEAKED REFERENCES

From a paper's bibliography to bibliometric dashboards, the path is long for references to be counted. Different actors using various deception techniques can sneak undue references in along this path.

2.1 | Context: The DOI and metadata registration process

As Figure 1 shows, after acceptance and before publication of papers, publishers register DOIs with Registration Agencies. One of the main agencies is Crossref, which enables registration of DOIs for a fee and hosts the publishers' metadata which is then made publicly available (Hendricks et al., 2020). Most publishers submit the reference lists of their papers to Crossref as part of the registered metadata (Singh Chawla, 2022).[2] Crossref is then used as a source by multiple platforms such as SpringerLink,[3] The Lens (Penfold, 2020), and Dimensions (Herzog et al., 2020).[4] Bibliometric platforms retrieve the metadata registered with Crossref inter alia to report indicators at the individual/institutional/journal levels, such as citation counts, impact factors, and *h*-indices.

2.2 | The manipulation … explained

Crossref makes available the metadata sent by their members, namely the publishers:

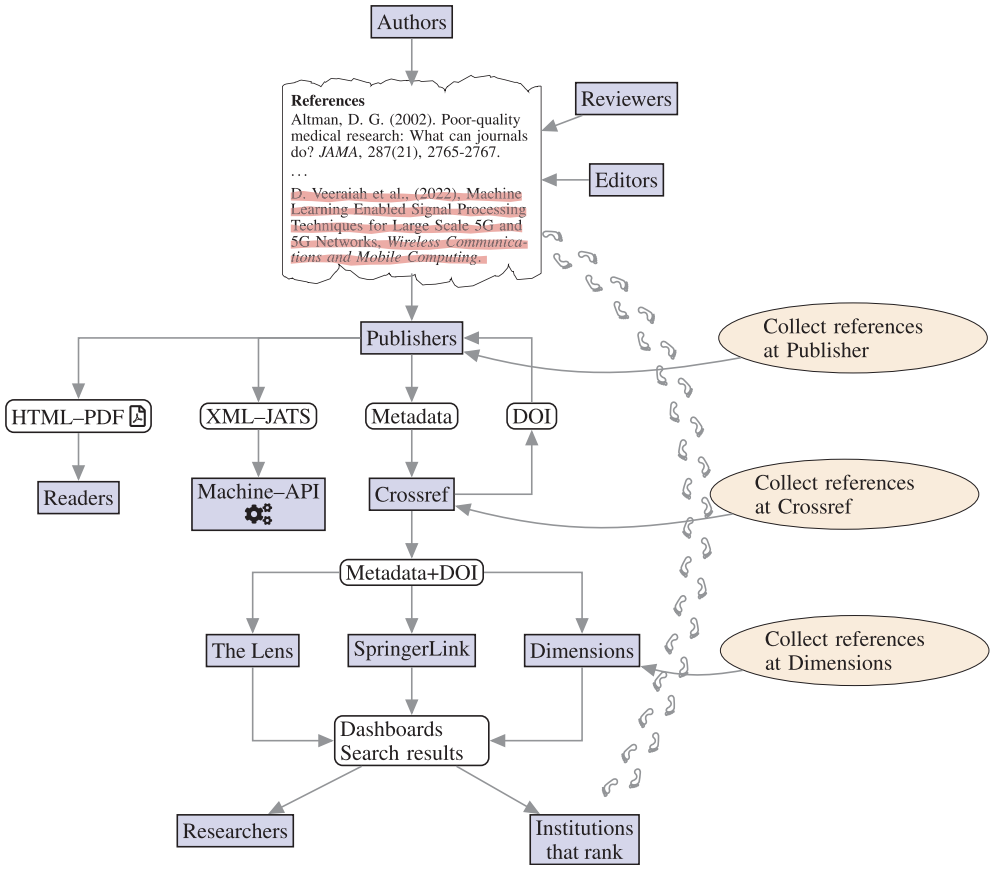

FIGURE 1 References' long path from authors to bibliometric dashboards: after Editorial and Peer-Review assessment, metadata are registered to a DOI provider (here Crossref). Metadata are then retrieved by bibliometric platforms (The Lens, SpringerLink, Dimensions) that provide various services, such as a search engine and bibliometric dashboards for institutions. A sneaked reference from our analysis (see Figure 2) appears highlighted in red, the different actors involved are in blue square, and our data collection process is shown in light orange ellipses.

> "Our metadata is provided to us by our members, and we don't curate or clean up the metadata in any way. We do insert metadata into outputs such as DOI matches for citations, recursive relationships, and clearly flag those pieces as being inserted by Crossref in our metadata outputs.
>
> This means, good or bad, metadata accuracy depends on the quality of metadata provided by our members."
>
> —'Metadata principles and practices' from Crossref
>
> https://www.crossref.org/documentation/principles-practices/

When registering a new publication and its references with Crossref, a publisher may sneak extra undue references in the metadata sent in addition to the ones originally present. Then, digital libraries (e.g., SpringerLink) and bibliometric platforms (e.g., Dimensions) harvest the fabricated metadata, including the undue citations. These sneaked references are processed and counted even if they are not present in the original publication.

This new way to manipulate citation counts relies on metadata manipulations which leave the original text untouched. This manipulation is made possible because Crossref trusts publishers to extract, report, and send them metadata about the publications, including the references. This trust is bound under their membership terms which include keeping metadata accurate and up-to-date. Crossref membership may be terminated for "fraudulent use of Identifiers or Metadata."[5] Effectively, because Crossref is not checking the accuracy of the metadata provided by publishers, this creates a "breach" within the information flow. The next section shows that this manipulation is actually in use by at least one of Crossref's 20,000 members.

3 | CASE STUDY: EVIDENCE OF SNEAKED REFERENCES IN THREE JOURNALS OF A GIVEN PUBLISHER

To provide evidence of citation counts manipulation, one needs to collect samples of metadata in three different places along the reference registration path depicted in Figure 1. Sneaked references are revealed when comparing the reference lists of publications as provided (1) by

FIGURE 2 PubPeer post https://pubpeer.com/publications/A172115FC8D0A5F44B31A18B08BB26 reporting a Hindawi journal article with more citations than downloads. Most citations appear not to match any of the references in the allegedly citing publications. After careful examination, it appeared that these were sneaked references: existing in the metadata only and not in the PDFs of the allegedly "citing" publications.

the publisher on its website, (2) on the metadata registry at Crossref, and (3) by a bibliometric platform: Dimensions.

As proof of the "sneaked references" manipulation happening, let us analyze three journals published by *Technoscience Academy*,[6] an Indian open access publisher and Crossref member. These three journals were selected after we identified incoherent metadata that we flagged in May 2022 on PubPeer (Figure 2). This case involves a Hindawi journal article published on 22 March 2022 and retracted as of 17 June 2023, due to a "systematic manipulation of the publication process." The Hindawi website showed a large number of citations ($n = 107$) for a publication that had been online for less than 2 months. On the screenshot in Figure 2, the number 107 stems from Altmetric, a service offered by Digital Science that sources data from publishers and Crossref.[7] Moreover, this number was far greater than the number of downloads ($n = 62$). These two observations combined had us suspect manipulations.

Further examination revealed that this Hindawi publication had no citations on Google Scholar. According to Dimensions, citations stemmed mostly from three main journals with 1000+ DOIs registered with Crossref. After careful verification, the PDF's citing publications did not contain any references to the Hindawi article. This is clear evidence that some references registered with Crossref as metadata (for the citing publication) do not exist in the full-text version of record. We assessed the extent of the discrepancy between the bibliographies of

(1) the published papers and (2) the metadata that were registered, hypothesizing that these two sets of references should be identical—except for undue sneaked references.

3.1 | Method to assess the extent of sneaked references

This section introduces a two-step method to measure differences between reference lists. First, we collect metadata about a publisher's catalog from three sources: the publisher's website, Crossref, and Dimensions. Second, we compare the reference lists as they appear in these three sources. We illustrate this method with the three largest journals published by *Technoscience Academy* and report numbers as of January 2023.

3.1.1 | Collecting metadata from Crossref

Crossref makes available the list of DOIs they have by journal and by publisher in the Crossref system.[8] For example, here are the DOIs of the journals registered for *Technoscience Academy*:

- 1063 DOIs minted for *IJSRSET*: the *International Journal of Scientific Research in Science, Engineering and Technology* at https://data.crossref.org/depositorreport?pubid=J325422.
- 1347 DOIs minted for *IJSRCSEIT*: the *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* at https://data.crossref.org/depositorreport?pubid=J326368.
- 1276 DOIs minted for *IJSRST*: the *International Journal of Scientific Research in Science and Technology* at https://data.crossref.org/depositorreport?pubid=J325454.

We retrieved the reference list of each publication by querying the Crossref API. For instance, https://api.crossref.org/works/10.32628/IJSRST229212 provides the metadata of publication https://doi.org/10.32628/IJSRST229212, including an attribute called `reference-count`. For this particular example, Crossref provided a list of 47 references (Figure 3).

3.1.2 | Metadata collection from the Publisher's web site

We retrieved the reference list of each publication identified in the previous section. Without any available API to retrieve metadata from *Technoscience Academy*, this step is specific to each journal. The journal articles from this publisher are available open access, in both PDF and HTML formats. We assumed that the reference lists provided in HTML conformed to the ones present in the PDF files—and verified this by visual inspection of a dozen cases. HTML pages feature a tab with the list of references that we collected via ad hoc scripts.

Our running example (https://doi.org/10.32628/IJSRST229212) has seven references present in the PDF and on the HTML page (Figure 4). The references present in HTML are also found in Crossref. But an additional set of 40 well-formed references turn out to be undue references to unrelated publications. This set comprises sneaked references that might have been added at registration time, or added subsequently in metadata updates at any time by the publisher.

3.1.3 | Metadata collection from Dimensions

Dimensions provides registered accounts for free, allowing users to query their database and export results up to 5 k publication records. We used the "Publisher" filter of Dimensions to collect the metadata of all papers published by *Technoscience Academy* and exported results using the "Export for bibliometric mapping" feature. The export came as a CSV file of 3634 publication records. One of the columns contains the reference list for each paper, as recorded by Dimensions.

According to this file, the article of the running example (https://doi.org/10.32628/IJSRST229212) has 13 references... to be compared to seven in HTML and 47 registered with Crossref. Visual inspection of the references found in Dimensions (Figure 3) reveals that none of these 13 references are from the original set of seven references (PDF and HTML; Figure 4).

Along the registration process, the seven original references were replaced by 13 undue sneaked references. The original version of the publication lists seven references; it was registered with Crossref with 40 undue sneaked references. Finally, Dimensions reports 13 references for this paper, all sneaked in. The seven original references appearing in the HTML/PDF as well as 27 sneaked references got lost along the path.

3.1.4 | Detecting sneaked and lost references

Tracing the propagation of individual references from one platform to another proved quite challenging due to the variability of reference formatting (e.g., APA, MLA, Chicago...). We decided to examine and compare the

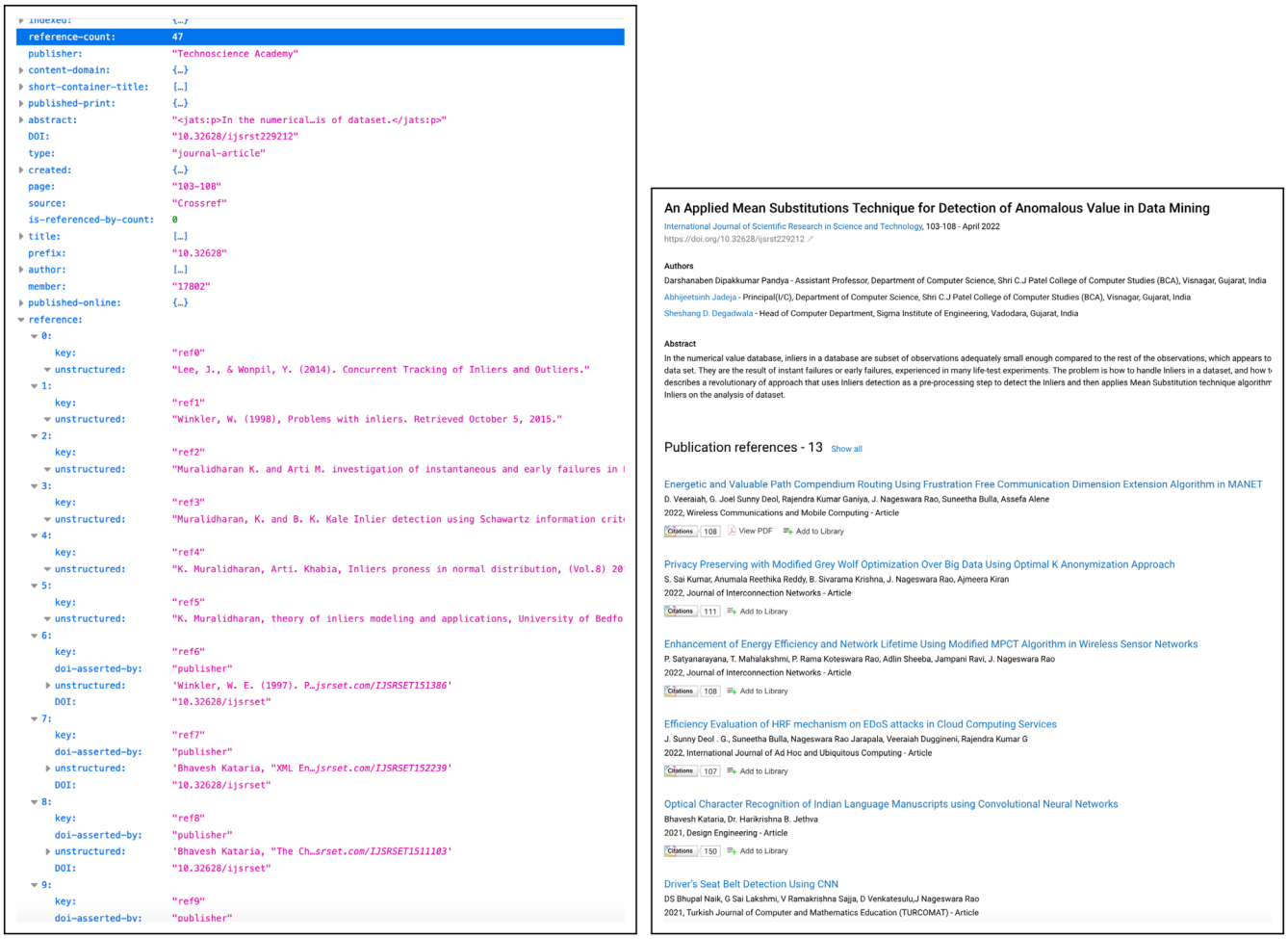

FIGURE 3 Reference list for publication https://doi.org/10.32628/IJSRST229212 as registered at Crossref (left: https://api.crossref.org/works/10.32628/IJSRST229212) and as retrieved from Dimensions (right: https://app.dimensions.ai/details/publication/pub.1146638907). Crossref provides the attribute `reference-count` (highlighted in blue) and a reference list of 47 references (numbers 0 to 9 shown). References 6 to 46 are sneaked references. Dimensions lists 13 references; none of them appear in the original paper (Figure 4).

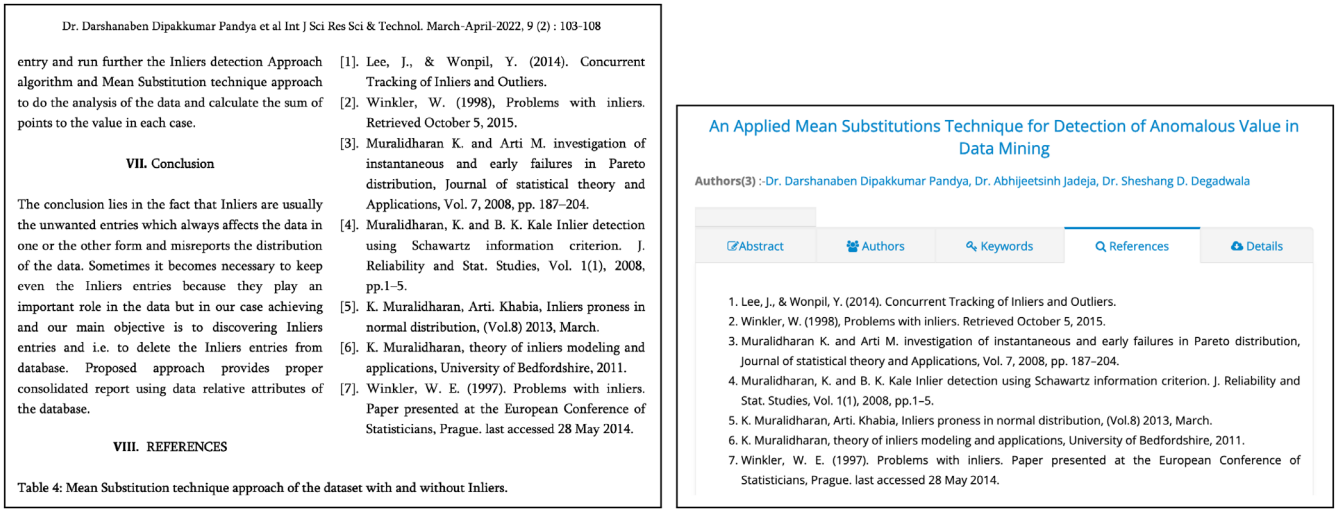

FIGURE 4 Reference list in PDF (left) and in HTML (right) versions of https://doi.org/10.32628/IJSRST229212. In this case, the PDF and HTML versions match each other, which is expected.

number of references to estimate inconsistencies between the size of the reference list in HTML/PDF versions and the registered metadata.

For each publication p, let R_C^p (resp. R_D^p) be the number of references registered with Crossref (respectively Dimensions) and S^p the number of references present in the PDF or HTML versions. Then $\delta_x^p = R_x^p - S^p$ given $x \in \{C, D\}$ estimates inconsistencies. The value δ_D^p (respectively δ_C^p) reflects inconsistencies between registered references with Crossref (respectively Dimensions) and those present in the HTML/PDF for publication p. Let us interpret δ_x^p:

- A zero value for δ_x^p indicates that, for publication p, the number of references registered in x equals the number of references listed in its PDF/HTML version. However, $\delta_x^p = 0$ does not guarantee that the registered references are the same as the references in the PDF/HTML.
- $\delta_x^p < 0$ reveals *lost* references: some are present in the publication p but are not registered. In that case δ_x^p is a lower bound of *lost* references.
- $\delta_x^p > 0$ is the lower bound of the number of *sneaked* references for publication p.

Let us illustrate the "lower bound" nuance on the running example: $p = \text{IJSRST229212}$. The number of sneaked references is underestimated when computing $\delta_D^p = R_D^p - S^p = 13 - 7 = 6$ in comparison with the exact number of sneaked references which is equal to 13 (see Figures 3 and 4). In this example, since $\delta_D^p > 0$ we cannot conclude that references are lost. However, comparing the content of the reference list allows us to see that all seven references of the HTML/PDF version are lost (see Figures 3 and 4). We can therefore see that δ_p^i also underestimates the number of lost references.

For a particular set $\mathcal{A}$ of journal articles, three publication subsets can be distinguished:

- The subset *OK* noted with O, contains publications for which $\delta_x^p = 0$.
- The subset *Sneaked* noted with S, contains publications for which $\delta_x^p > 0$, where we have evidence that references have been sneaked in.
- The subset *Missing* noted with M, contains publications for which $\delta_x^p < 0$, where we have evidence that references are lost.

For a set $\mathcal{A}$, we can compute Δ_x^S (respectively Δ_x^M) the overall lower bound of sneaked (respectively lost) references with the sum over $p \in \mathcal{A}$ of positive (respectively negative) δ^i:

$$\Delta_x^S = \sum_{p \in S} \delta_x^p,$$

$$\Delta_x^M = \sum_{p \in M} \delta_x^p.$$

It is also possible to see if references found in publications of the *Sneaked* set benefit a few people or a few journals in particular. We detail the results of our analysis below.

3.2 | Results

3.2.1 | Quantitative analysis

The lower bound of sneaked (Δ_x^S) and lost references (Δ_x^M) for the set of journal articles from three journals presented previously are given in Tables 1 and 2. Data were collected from three different sources (publisher's website, Crossref, and Dimensions). Differences observed between the HTML/PDF and Crossref (Δ_C^x) are shown in Table 1, whereas Table 2 shows the differences between HTML/PDF and Dimensions (Δ_D^x).

In Table 1 an article is counted in the *Sneaked* set if the reference list in the HTML/PDF is shorter than the one found at Crossref ($\delta_C^i > 0$). Among the 3506 articles published by these three journals, at least 230 articles contain more references than they should. $\Delta_C^S = 5978$ is the lower estimation of the total number of references that were unduly sneaked in when the metadata was deposited with Crossref. This represents an augmentation of 9.8% of the original set of references (60, 635). Out of 65,836 references that were registered, $9.1\% = 5978/65{,}836$ are therefore *Sneaked in*. In addition, for 73 articles some references were missing (status *Missing*), and in total, at least 777 references are missing in Crossref. This represents a decrease of $1.2\% = 777/60{,}635$.

Table 2 compares the sizes of the reference lists in the HTML/PDF and in Dimensions. For the vast majority of publications some references are missing. This is the case for 3184 articles (status *Missing*) out of the total of 3506. For these publications, some references can be seen in the HTML/PDF but are not registered in Dimensions. In total, at least $40.7\% = 24{,}712/60{,}635$ of the original references are missing in Dimensions. For 120 publications, more references can be found in Dimensions than in the HTML version (status *Sneaked*). In total, at least $2.7\% = 10163/6939$ of references registered for these journals are undue sneaked references.

Status	Number of articles		Number of references		
			In Crossref	In HTML	In Crossref–In HTML
OK	O	3203	55,252	55,252	0
Sneaked	S	230	10,404	4426	$\Delta_C^S = 5978$
Missing	M	73	180	957	$\Delta_C^M = -777$
Total	$\mathcal{A}$	3506	65,836	60,635	

TABLE 1 Statistics on the *Technoscience Academy* corpus showing the discrepancies between the references found in the versions of record (HTML/PDF) and the ones registered at Crossref.

TABLE 2 Statistics on the *Technoscience Academy* corpus showing the discrepancies between the references found in versions of record (HTML/PDF) and the ones registered in Dimensions.

Status	Number of articles		Number of references		
			In Dimensions	In HTML	In Dimensions–In HTML
OK	O	202	2414	2414	0
Sneaked	S	120	2672	1656	$\Delta_D^S = 1016$
Missing	M	3184	31,853	56,565	$\Delta_D^M = -24,712$
Total	$\mathcal{A}$	3506	36,939	60,635	

3.2.2 | Qualitative analysis

To understand the discrepancies highlighted above, we decided to closely inspect some examples of problematic cases. In particular, we decided first to inspect the cases displaying significantly large discrepancies. For instance:

- https://doi.org/10.32628/ijsrset21852 has 150 references in its HTML version but 300 are registered with Crossref. We noticed that the reference list is duplicated. Only 114 references can be found in Dimensions. Among the $186 = 300 - 114$ missing references, an example is a reference claimed to be a technical report from the Liverpool John Moores University, UK by Younis & Kifayat which, after verification, is not indexed by Dimensions (but is indexed in Google Scholar).
- https://doi.org/10.32628/ijsrst229394 lists 27 references in HTML/PDF but $108 = 4 \times 27$ were registered in Crossref. We noticed that the same set of 27 references were registered four times. Nevertheless, only 19 references can be found in Dimensions such that eight references are missing.

From these examples, we can conclude that lost references (status *Missing*) may often result from a failure to attach a given reference to a *citable item* because of incomplete or erroneous registered metadata with Crossref. It is important to note that some types of references are, by definition, not indexed in Dimensions (private correspondences, songs...). We can also conclude that some of the *sneaked* references may be due to the careless management of metadata by the publisher, resulting in such erroneous registrations. These duplications however do not seem to propagate to Dimensions: at most one occurrence of the duplicated references was present.

However, not all *sneaked* references can be explained by careless metadata registration as can be seen in the following example. The article https://doi.org/10.32628/ijsrst229154 has an HTML/PDF version which lists 23 references. However, 63 can be found in Crossref and 33 in Dimensions. An analysis of the 10 *sneaked* references in Dimensions reveals that they benefit mainly two authors (Initials JNR & BK). Therefore it seems that additional references may be *sneaked in* to benefit specific scholars. To verify this hypothesis, we computed the most frequent words in Crossref's metadata (`reference` field, see Figure 3) for papers identified as containing *sneaked* references. This analysis reveals that undue *sneaked* references mostly benefited two scholars and a few journals published by *Technoscience Academy*:

- One person with initials JNR benefited from 3103 extra citations.
- One person with initials BK benefited from 1564 extra citations.
- The *International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET)* gained 826 extra citations.
- The *International Journal of Advanced Science and Technology (IJAST)* was unduly cited 537 times.
- The *Turkish Journal of Physiotherapy and Rehabilitation* appeared 428 times in sneaked references.

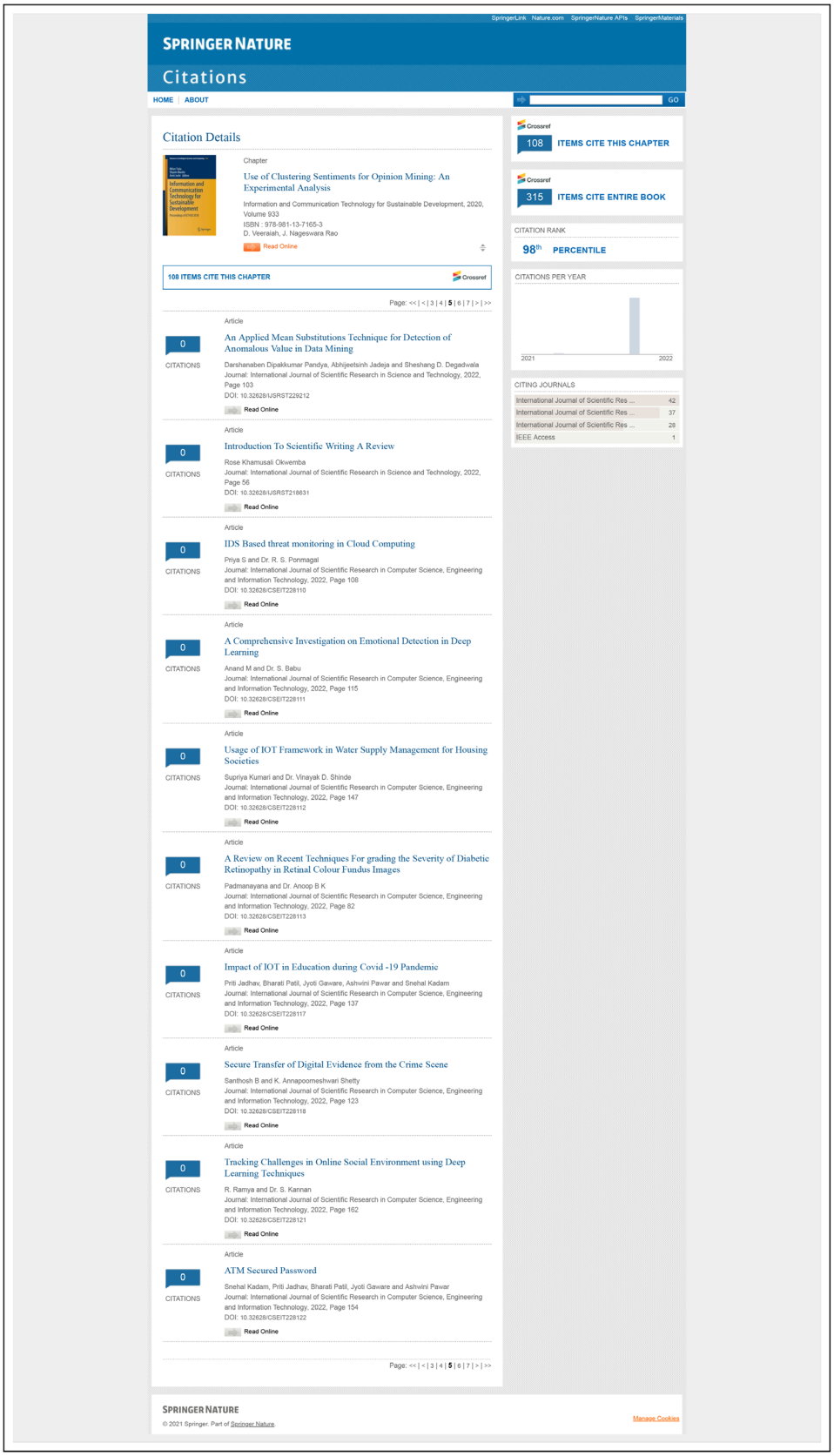

FIGURE 5 Undue citations received by a Springer article coauthored by JNR. Screenshot of SpringerLink https://citations.springernature.com/item?doi=10.1007/978-981-13-7166-0_62 as of 6 January 2023. Some of the 108 citations are due to sneaked references: See, for instance, the top-listed reference that is our running example: https://doi.org/10.32628/IJSRST229212 (Figure 4).

Abalkina et al. (2022) identified as "hijacked" these last two journals in the list above. Furthermore, it is worth noting that individual JNR coauthored the article in Figure 2. Individual BK is the Executive Editor of Technoscience Academy's *IJSRCSEIT* (*International Journal of Scientific Research in Computer Science,*

Engineering and Information Technology[9]) and registered its internet domain.[10] Finally, let us use our running example again to show how sneaked references propagated from Crossref to other bibliometric platforms, such as SpringerLink (Figure 5).

3.3 | External validation of our findings

We initially posted our findings as a preprint at the time of submission of this manuscript (Besançon et al., 2023). Following this, we informed both Crossref and Dimensions of our findings and linked to our preprint. Dimensions, on the one hand, acknowledged that the issue exists and explained that they base their data on the one they obtain from Crossref. On the other hand, Crossref did confirm to us the presence of references in the metadata that are not in the full-text of the published manuscripts. In an interview for *Retraction Watch* with Singh Chawla (2023), Crossref confirmed they had not seen or found such issues in their repository before; it was the first time they had heard about references in the metadata that were not in the articles. Following this, Crossref conducted an investigation and asked the publisher to explain the discrepancies between (1) the bibliographies of the published articles and (2) the bibliographic metadata the publisher had registered at Crossref. The publisher took action and mostly fixed the problem: now the Crossref metadata reflect the bibliographies present in the articles.

4 | DISCUSSIONS: OUTCOMES AND POSSIBLE COUNTERMEASURES

Crossref, the largest DOI registration agency, provides metadata to many downstream users, such as Dimensions, The Lens, and SpringerLink. The numbers provided by Crossref, and these downstream services guide funding decisions and state policies. Our results shed light on flawed metadata affecting reference registration and, in turn, citation counts. We have identified a new source of quality problems: undue references *sneaked in* when the metadata are registered. To the best of our knowledge, the vulnerability we discovered is the first documented manipulation of metadata that does not modify the underlying PDF/HTML article. Our analysis highlights that the problems may arise for different reasons, ranging from publishers' careless management of metadata to potential citation counts manipulations. We indeed observed artificially inflated citation counts that seem to mostly benefit specific scholars or scientific journals. The metadata registration process is vulnerable: it was and is likely to be abused by various actors (authors, journals, publishers) to unduly inflate their citation counts. Additionally, this vulnerability, if exploited, may hinder other scholars who will not obtain their deserved citations.

To prevent the use of this vulnerability affecting the computation of citation counts, many actions and countermeasures exist. The most trivial ones imply the three key actors (see Figure 1) checking each others' metadata:

- Publishers and Crossref should check and compare the coherence of references registered and the ones actually present in publications (PDF/HTML).
- Bibliometric platforms and Crossref should check on each other to make sure that citation counts are coherent with registered metadata.
- Bibliometric platforms and publishers should check on each other, to ensure that citations credited to articles are indeed supported by the associated references in the citing publications.

A more extensive countermeasure would involve third parties independently auditing the whole process: from checking the metadata uploaded into metadata registration agencies to checking the validity of citation counts. Thanks to the Initiative for Open Citations (https://i4oc.org), and a Crossref board vote in March 2022, 100% of the references publishers deposit with Crossref are publicly available (Schiermeier, 2017; Shotton, 2013; Singh Chawla, 2022). However, only 52% (54.8 M/105.0 M) of the registered journal articles have references deposited.[11] Open and free access to APIs and references at various steps of the process illustrated in Figure 1 is required to enable third parties to check the global quality of the provided data. Once inaccuracies are detected, corrective actions as well as potential sanctions must take place. Crossref, Elsevier's Scopus, and Clarivate's Web of Science currently have the ability to act against offending publishers and journals.[12] We suggest an extension to the COPE (2019) "discussion document [that] defines the key issues and existing solutions around unethical citation practices" (1) to account for deceitful *sneaked references* and (2) to specify the appropriate reporting and editorial actions needed. The publishing community must give extra attention to correct erroneous reference metadata in addition to correcting the scholarly literature in due time (Besançon et al., 2022).

5 | CONCLUSION

This article showed evidence of an undocumented vulnerability affecting the process of metadata registration for academic works. Despite being absent from the Version of Record (in the HTML/PDF), sneaked references exist in the metadata, which in turn inflates citation counts unduly. The method we proposed estimates lower bounds for the number of references that were lost and sneaked in. Through a case study, we show that this vulnerability is actually exploited. One still needs to apply this method on the entire literature to estimate the extent of the "sneaked/lost references" issue at the global scale. This proves challenging as most of the global citation graph is inaccessible or accessible but presented in heterogeneous forms.

Our work questions the quality and veracity of the reference metadata harvested in Crossref and used by bibliometric platforms, such as Dimensions. These metadata support commercial bibliometric services and inform influential rankings of institutions and individuals. All actors involved should be held accountable for the quality of the data they provide, share, and sometimes trade. We believe they must prevent metadata distortion, keeping in mind the inerrant drawbacks of the extensive use of citation metrics, fueling elaborate cheating schemes.

ACKNOWLEDGMENTS

We thank Dr. Nick H. Wise for discussions and feedback on the first version of this manuscript. We also wish to thank the reviewers of this article for their insightful feedback. We are indebted to Ginny Hendricks and Fabienne Michaud from Crossref for their valuable comments on the revised manuscript. CL and GC acknowledge the NanoBubbles project that has received Synergy grant funding from the European Research Council (ERC), within the European Union's Horizon 2020 program, grant agreement no. 951393 (https://doi.org/10.3030/951393).

DATA AVAILABILITY STATEMENT

We release Supporting Information for reproducibility purposes and future scientific literature screening. The code developed to collect and analyze the data reported in this article is archived at Zenodo (https://doi.org/10.5281/zenodo.8388930).

ORCID

Lonni Besançon https://orcid.org/0000-0002-7207-1276
Guillaume Cabanac https://orcid.org/0000-0003-3060-6241
Cyril Labbé https://orcid.org/0000-0003-4855-7038
Alexander Magazinov https://orcid.org/0000-0002-9406-013X

ENDNOTES

[1] https://pubpeer.com/search?q=%22citation+plantation%22.

[2] https://www.crossref.org/documentation/schema-library/markup-guide-metadata-segments/references/.

[3] https://citations.springernature.com/about.

[4] To the best of our knowledge, Google Scholar relies on various sources and crawling methods (Van Noorden, 2014).

[5] See 9.a.iii.3 in https://www.crossref.org/membership/terms/.

[6] https://technoscienceacademy.com.

[7] Crossref reported 107 citations for this paper, see the attribute `is-referenced-by-count` shown at https://web.archive.org/web/202205/http://api.crossref.org/works/doi/10.1155/2022/3685419.

[8] https://www.crossref.org/06members/51depositor.html.

[9] See https://web.archive.org/web/202401/https://ijsrcseit.com/editorial.php.

[10] See https://web.archive.org/web/202401/https://who.is/whois/technoscienceacademy.com.

[11] See the Crossref queries: numerator https://api.crossref.org/works?filter=type:journal-article,has-references:true and denominator https://api.crossref.org/works?filter=type:journal-article.

[12] See https://www.crossref.org/operations-and-sustainability/membership-operations/revocation/, https://www.elsevier.com/products/scopus/content/content-policy-and-selection, and https://support.clarivate.com/ScientificandAcademicResearch/s/article/Journal-Citation-Reports-Explanation-of-Missing-Dropped-or-Suppressed-Journals.

REFERENCES

Abalkina, A., Cabanac, G., Labbé, C., & Magazinov, A. (2022). Improper legitimization of hijacked journals through citations. *arXiv*. (Oral presentation at PRC'22, the 9th international congress on peer review and scientific publication, https://peerreviewcongress.org/?p=1871) https://doi.org/10.48550/arXiv.2209.04703

Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, highquality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies*, *1*(1), 377–386. https://doi.org/10.1162/qss_a_00019

Baccini, A., De Nicolao, G., & Petrovich, E. (2019). Citation gaming induced by bibliometric evaluation: A country-level comparative analysis. *PLoS One*, *14*(9), e0221212. https://doi.org/10.1371/journal.pone.0221212

Beel, J., & Gipp, B. (2010). On the robustness of Google Scholar against spam. In *HT'10: Proceedings of the 21st ACM conference on hypertext and hypermedia* (pp. 297–298). ACM. https://doi.org/10.1145/1810617.1810683

Besançon, L., Bik, E., Heathers, J., & Meyerowitz-Katz, G. (2022). Correction of scientific literature: too little, too late! *PLoS Biology*, *20*(3), e3001572. https://doi.org/10.1371/journal.pbio.3001572

Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2023). Sneaked references: Cooked reference metadata inflate citation counts. *arXiv*. https://doi.org/10.48550/arXiv.2310.02192

COPE. (2019). Citation manipulation. https://doi.org/10.24318/cope.2019.3.1

Crous, C. J. (2019). The darker side of quantitative academic performance metrics. *South African Journal of Science*, *115*(7/8), 1–3. https://doi.org/10.17159/sajs.2019/5785

Davis, P. (2016). Visualizing citation cartels. Retrieved from https://wp.me/peaj1R-cdk (Scholarly Kitchen)

Delgado López-Cózar, E., Robinson-García, N., & Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, *65*(3), 446–454. https://doi.org/10.1002/asi.23056

Foley, J. A., & Valkonen, L. (2012). Are higher cited papers accepted faster for publication? [Editorial]. *Cortex*, *48*(6), 647–653. https://doi.org/10.1016/j.cortex.2012.03.018

Franck, G. (1999). Scientific communication—A vanity fair? [Essays on science and society]. *Science*, *286*(5437), 53–55. https://doi.org/10.1126/science.286.5437.53

Haley, M. R. (2017). On the inauspicious incentives of the scholar-level h-index: An economist's take on collusive and coercive citation. *Applied Economics Letters*, *24*(2), 85–89. https://doi.org/10.1080/13504851.2016.1164812

Heathers, J. A., & Grimes, D. R. (2022). Impact factor manipulation—the mechanics behind a precipitous rise in Impact Factor: A case study from the British Journal of Sports Medicine (OSF preprint). https://doi.org/10.17605/osf.io/4c6xa

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

Hinchliffe, L. J. (2022). The version of record as a central organizing concept in scholarly publishing. *Information Services & Use*, *42*(3–4), 309–314. https://doi.org/10.3233/isu-220164

Kojaku, S., Livan, G., & Masuda, N. (2021). Detecting anomalous citation groups in journal networks. *Scientific Reports*, *11*(1), 14524. https://doi.org/10.1038/s41598-021-93572-3

Labbé, C. (2010). Ike Antkare, one of the great stars in the scientific firmament. *ISSI Newsletter*, *6*(2), 48–52. Retrieved from https://www.issi-society.org/media/1126/newsletter22.pdf

Lawrence, P. A. (2007). The mismeasurement of science. *Current Biology*, *17*(15), R583–R585. https://doi.org/10.1016/j.cub.2007.06.014

Oransky, I. (2022, November 15). Why misconduct could keep scientists from earning Highly Cited Researcher designations, and how our database plays a part. Retrieved from https://retractionwatch.com/?p=126033 (Retraction Watch)

Penfold, R. (2020). Using the Lens database for staff publications. *Journal of the Medical Library Association*, *108*(2), 341–344. https://doi.org/10.5195/jmla.2020.918

Puebla, I., Polka, J., & Rieger, O. Y. (2021). Preprints: Their evolving role in science communication. MetaArXiv (preprint). https://doi.org/10.31222/osf.io/ezfsk

Schiermeier, Q. (2017). Initiative aims to break science's citation paywall [news]. *Nature*. https://doi.org/10.1038/nature.2017.21800

Shotton, D. (2013). Publishing: Open citations [comment]. *Nature*, *502*(7471), 295–297. https://doi.org/10.1038/502295a

Singh Chawla, D. (2022). Five-year campaign breaks science's citation paywall [news]. *Nature*. https://doi.org/10.1038/d41586-022-02926-y

Singh Chawla, D. (2023, October 9). How thousands of invisible citations sneak into papers and make for fake metrics. Retrieved from https://retractionwatch.com/?p=128012 (Retraction Watch)

Szomszor, M., Pendlebury, D. A., & Adams, J. (2020). How much is too much? The difference between research influence and self-citation excess. *Scientometrics*, *123*(2), 1119–1147. https://doi.org/10.1007/s11192-020-03417-5

Van Noorden, R. (2014). Google Scholar pioneer on search engine's future. *Nature*. https://doi.org/10.1038/nature.2014.16269

Van Noorden, R. (2020a). Highly cited researcher banned from journal board for citation abuse. *Nature*, *578*(7794), 200–202. https://doi.org/10.1038/d41586-020-00335-7

Van Noorden, R. (2020b). Signs of "citation hacking" flagged in scientific papers. *Nature*, *584*(7822), 508. https://doi.org/10.1038/d41586-020-02378-2

Wren, J. D., & Georgescu, C. (2022). Detecting anomalous referencing patterns in PubMed papers suggestive of author-centric reference list manipulation. *Scientometrics*, *127*(10), 5753–5771. https://doi.org/10.1007/s11192-022-04503-6

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Besançon, L., Cabanac, G., Labbé, C., & Magazinov, A. (2024). Sneaked references: Fabricated reference metadata distort citation counts. *Journal of the Association for Information Science and Technology*, 1–12. https://doi.org/10.1002/asi.24896