When Scientific Citations Go Rogue: Uncovering 'Sneaked References'

Lonni Besançon, Guillaume Cabanac

▶ **To cite this version:**

Lonni Besançon, Guillaume Cabanac. When Scientific Citations Go Rogue: Uncovering 'Sneaked References'. 2024. hal-04791003

HAL Id: hal-04791003
https://hal.science/hal-04791003v1

Submitted on 19 Nov 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

When Scientific Citations Go Rogue: Uncovering ‘Sneaked References’

Lonni Besançon, Guillaume Cabanac

The image of a researcher working alone, aside from the world and the rest of the wider scientific community, is just a classic yet misguided image. Research is, by essence, built on continuous exchange within the scientific community: first, to understand the work of others, and then to share one's own findings.

Reading and writing articles published in journals or presented at conferences are central and unavoidable activities for researchers. When writing an article, it is crucial to cite the work of peers to describe context, detail sources of inspiration, and explain differences in approaches and results. Being cited by other researchers for "good reasons" is often considered to be a key measure of the visibility of one's own work.

But what happens when this citation system is manipulated? A recent *JASIST* article [1] by our team of "academic sleuths" part of “Le Collège Invisible” reveals an insidious method to artificially inflate citation counts through metadata manipulations: "sneaked references."

The Hidden Manipulation

Scientific publications and their inherent functioning, including potential flaws and their causes, are becoming more and more common topics in science and popular science alike. Just last year more than 10,000 scientific articles were retracted. A large amount of the literature documents the issues around citation gaming and the harm it causes the scientific community and its credibility. Citations to scientific work abide by a standardised referencing system: each reference explicitly mentions at least the title, authors' names, publication year, journal or conference name, and page numbers of the cited publication. These details are stored as metadata, not visible in the article’s text directly, but assigned to a DOI (Digital Object Identifier), a unique identifier for each scientific publication.

References in a scientific publication allow authors to justify methodological choices or present the results of past studies, stressing the iterative and collaborative nature of science. However, we found, through an opportunistic encounter, that some unscrupulous actors have added extra references, invisible in the text but present in the article's metadata during submission to scientific databases. The result? Citation counts for certain researchers or journals skyrocket without valid reasons since these references were not cited by the authors in their articles.

A New Type of Fraud and an Opportunistic Discovery

The investigation began with Guillaume Cabanac, a professor at the University of Toulouse, who posted on PubPeer, a website dedicated to post-publication peer review, where scientists discuss and analyse publications. He noticed an inconsistency: a Hindawi article, likely fraudulent due to awkward phrases, had far more citations than downloads, which is very unusual. This post caught the attention of several "scientific sleuths" who are now the authors of the aforementioned *JASIST* article [1]. We tried to find articles citing the initial article using a scientific search engine. Google Scholar found none, while Crossref and Dimensions did. The difference? Google Scholar is very likely to mostly rely on the article's main text, whereas Crossref and Dimensions use metadata provided by publishers.

To understand the extent of the manipulation, we therefore decided to dig further and examined three scientific journals published by the same publisher as the questionable article:Technoscience Academy. Our investigation consisted in three steps:

1. **HTML/PDF Articles**: We listed the references explicitly present in the HTML or PDF versions.
2. **Crossref Metadata**: We compared these lists with the metadata recorded by Crossref, discovering extra references added in the metadata but not appearing in the articles.
3. **Dimensions**: We checked Dimensions, a bibliometric platform that uses Crossref as a metadata source, finding further inconsistencies.

In the journals published by Technoscience Academy, at least 9% of recorded references were "sneaked references." These additional references were only in the metadata, distorting citation counts and giving certain authors an unfair advantage. Some legitimate references were also "lost": not present in the metadata. In addition, when analysing the sneaked references, we found that they highly benefited some researchers. Our analysis revealed that a single researcher (associated with Technoscience Academy, by the way) benefited from more than 3000 additional illegitimate citations for instance, while some journals from the same publishers benefited from a couple of hundreds additional sneaked citations.

We finally wanted our results to be externally validated. We first posted our study as a preprint and informed both Crossref and Dimensions of our findings and gave them a link to the preprinted investigation we did. Dimensions acknowledged the issue exists and confirmed that part of their database reflects Crossref's data. Crossref also confirmed in Retraction Watch the issue and highlighted that this was the first time that they were notified of such a problem in their database. The publisher, based on Crossref's investigation, has now taken action to fix the problem.

Implications and Potential Solutions

Why is this discovery important? Citation counts heavily influence research funding, academic promotions, and institutional rankings. Manipulating citations can lead to unjust decisions based on false data. More worryingly, this discovery raises questions about the integrity of scientific impact measurement systems, a concern highlighted by researchers for years. These systems can be manipulated to foster unhealthy competition among researchers, tempting them to take shortcuts to publish faster or achieve more citations.

To combat this practice, we therefore suggest several measures:

- Rigorous verification of metadata by publishers and agencies like Crossref.
- Independent audits to ensure data reliability.
- Increased transparency in managing references and citations.

This study is the first, to our knowledge, to report a manipulation of metadata and discusses the impact this may have on the evaluation of researchers. It may also be that the study highlights, yet again, that the overreliance on metrics to evaluate researchers, their work and their impact may be inherently flawed and wrong. Not only is it likely to promote questionable research practices (e.g., HARKing [3], data manipulation, plagiarism, salami-slicing of papers, …) it also does not promote a greater transparency which would be the key to more robust [4] and efficient [5] research. Although the problematic citation metadata and sneaked references have now been apparently fixed, the correction may have, as it is often the case for scientific correction [6], happened too late.

[1] 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

[2] Van Noorden, R. (2023). More than 10,000 research papers were retracted in 2023—a new record. Nature, 624(7992), 479-481. https://doi.org/10.1038/d41586-023-03974-8

[3] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 141, 1–12. https://doi.org/10.1145/3173574.3173715

[4] Besançon, L., Peiffer-Smadja, N., Segalas, C. et al. Open science saves lives: lessons from the COVID-19 pandemic. BMC Med Res Methodol 21, 117 (2021). https://doi.org/10.1186/s12874-021-01304-y

[5] Chalmers, Iain et al. Avoidable waste in the production and reporting of research evidence. The Lancet, Volume 374, Issue 9683, 86 - 89. https://doi.org/10.1016/S0140-6736(09)60329-9

[6] 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