Retracted Papers:
the Good, the Bad, and the Ugly

Ethan Haque

May 7, 2024

An unsettling amount of published scientific literature is provably wrong. Since 2014, the number of papers flagged for being fundamentally flawed to the extent their results are unreliable has steadily grown with over ten thousand flaggings in 2023 [11]. This flagging mechanism is called retraction. Retracted papers are not removed from the scientific literature, but are instead marked as retracted and often accompanied with a notice written by the authors explaining the reason behind the retraction [1]. To the reader, a retraction sometimes appears as a physical watermark on an electronic copy of a paper. Other times, a retraction notice may appear on the website where the paper is distributed. Sometimes a reader has no indication that a paper has been retracted at all. This study is concerned with how the general population interacts with retracted scientific literature. In particular, it seeks to understand how retracted papers affect information in textbooks and how conclusions from retracted papers can enter the classroom. It uses the Retraction Watch Database and the analysis of citation networks to accomplish this [10].

A natural question to ask about retracted papers is how can retraction happen in the first place? Isn't the peer-review process and, more broadly, the scientific method supposed to stop these papers from getting published? An incomplete answer is in an ideal world this would not happen, but researchers make mistakes. Scientific research is naturally self-correcting so these mistakes are eventually found then corrected. This is missing a more sinister aspect of retracted papers: deliberate academic fraud. Unfortunately, academic fraud is becoming a more common reason for retraction. One study found that, ”the number of papers retracted for fraud increased more than seven-fold in the 6 years between 2004 and 2009. During the same period, the number of papers retracted for a scientific mistake did not even double” [9]. This is not to say that academic fraud is committed more often now than it has in the past, but that we are getting better at catching it. It is likely this observed increase in academic fraud cases is due to better review methods and more scrutiny by publishing journals over the past two decades [9].

Part of the reason academic fraud is so common is the incentives for scientists are not aligned with the goals of the scientific method. Research is more likely to be published if it produces positive results. This creates what is called publication bias, where studies with negative or inconclusive results are less likely to be published, even though they are also important for the advancement of scientific knowledge. Consequently, researchers might be tempted to manipulate data or results to make them appear more favorable. Moreover, the pressure to publish frequently and in prestigious journals—often summed up in the phrase ”publish or perish”—further exacerbates this issue. Such pressures can lead to cutting corners, fabrication, or plagiarism [2].

Retracted papers do not exist in a vacuum. They have tangible effects in public policy, business decisions, health, and education. Take for example the 1998 paper authored by Andrew Wakefield et al. fraudulently linking the measles, mumps, and rubella (MMR) vaccine and autism [12]. This paper alone has had devastating public health consequences spanning decades [7, 4, 6]. Many studies examine the relationship between retracted research like Wakefield's and how the public approaches topics like health and environmental policy. Many of these studies emphasize the need for more comprehensive education around misinformation and scientific literacy [6, 5, 8, 3]. This begs another question, though. What retracted scientific research is taught in the classroom?

Recently, new databases tracking retracted research and their connections to other papers have made it possible to examine the spread of retracted research into new domains like public policy and education [4]. The remainder of this paper details a computational methodology to track the flow of retracted research into K-12 textbooks, and describes an interface where this information may be disseminated.

The citation is the main unit of analysis in this study. The goal is to generate a citation graph with directed paths from K-12 textbooks to retracted papers. These paths indicate specific instances where misinformation may have entered a textbook. Each node in this graph is a unique work, and each directed edge going from X to Y indicates X cites Y. This graph must be a directed acyclic graph (DAG). It must be a DAG because a cycle would indicate a work citing itself, directly or indirectly, which is nonsense. To see this more concretely, consider Figure 1, where nodes p_1, p_2, and p_3 represent different papers. p_1

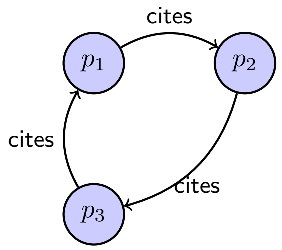

Figure 1: In this graph, there is a directed cycle implying each paper eventually cites itself.

cites p_2, which cites p_3, which then cites p_1. This circular referencing implies that the information flows back to its origin and a work is both a source and a consumer of its own information.

This graph should also ideally have paths from textbooks to retracted papers no longer than a threshold value μ. As the lengths of paths in this graph increase, the relationship between the original textbook and retracted paper become weak and hard to reason about. It's not clear what value μ should take, but an initial guess for a good value is $\mu = 3$. This allows for at most one intermediate paper between a textbook and retracted paper.

Building this graph requires three data sources. First, the Retraction Watch Database provides the source for retracted papers [10]. It provides several key pieces of information including the DOI of the paper, authors, title, reasons for retraction, and date of retraction. As of May 2024 it contains over $48,000$ retracted papers. Second, the contents of the textbooks would need to be acquired. Each state's Department of Education usually oversees the adoption of textbooks for public schools. Many states have an official list of approved textbooks and instructional materials for each subject. As a last resort, this data can be obtained using the the Freedom of Information Act. The contents of these textbooks may be expensive to acquire, and downstream publication must respect fair use. Third, the citation network for the works cited in both the textbooks and the retracted papers must be built. This can be done using a service like Altmetric. With these three pieces of information, it is possible to construct a DAG where a path represents works that build upon one another.

Linking the citations together will be a challenge. Without an explicit unique identifier like a DOI approximate matching techniques will be necessary. This matching step is crucial to get right since the analysis of this graph will require the connections to be accurate. At a large enough scale it will become impossible to verify the accuracy of each connection. Uncertain connections should be removed and not included in analysis. This will miss possible connections between textbooks and retracted papers, but this is better than introducing a connection that should not exist and creating a false positive connection. Parsing the citations themselves can also be difficult. If the citations are stored in a format like PDF then several tools like cloud text extraction services, pdf manipulation libraries, and custom OCR/layout analysis software may be needed.

Once, constructed, a standard data science technology stack like Python, NumPy, Pandas, and graph-tool will suffice to analyze this graph. This graph can answer several interesting questions by itself like: how often do textbooks cite retracted papers compared to non-retracted research. Does this vary with respect to the subject of the textbook? How have retracted citation trends changed over time adjusting for the increasing rate of paper retraction? Are intermediate papers more common to cite than retracted papers themselves?

A deeper, more nuanced understanding of how these citations affect the content of textbooks requires close reading techniques instead of a distant reading approach. Randomly sampling several paths between textbooks and retracted papers, then investigating how the content is used in the context of the textbook will reveal patterns of how misinformation is treated in these textbooks. Some textbooks might present flawed studies as accurate, whereas others may reference retracted works to highlight past scientific controversies or discredited theories.

The resulting graph for this study can be made freely available online including metadata like the authors of the papers and textbooks, dates, titles, etc. However, the actual content of most of these works is likely protected and cannot be disseminated. Specific instances of misinformation and transformative examples using direct quotes from the text can be published under fair use.

Aggregate statistics are also OK to publish.

Crowdsourcing additional examples and adding to the citation network could make this work much more interactive. Since the graph can be made freely available readers may investigate paths from retracted papers to textbooks on their own—provided they can find their own access to the source materials. A user may be able to add extra nodes and paths to the graph and add annotations to the paths between textbooks and retracted works detailing the significance of the path. This collaborative, crowdsourced approach would enhance the graph's comprehensiveness and accuracy, as individual contributions from educators, researchers, and concerned readers could help fill gaps and provide a broader understanding of how misinformation spreads through citation networks. Annotations might include notes on the nature of the misinformation, possible reasons why a particular retracted paper was cited, and how the content was interpreted or presented in the textbooks. The existing information provided in the graph can be an example for the crowdsourced information. Additional tutorial resources and detailed documentation would be needed as well.

This is a difficult process to get right. It will require careful treatment to avoid false positive links as much as possible. However, the value of this work lies in shedding light on how retracted research affects educational content. By understanding the paths through which misinformation reaches textbooks, educators and policymakers can address these issues more effectively and ensure students are learning from accurate and reliable sources. Collaboration will be key, and the crowdsourced model will empower a diverse range of voices to contribute to a more comprehensive picture.

References

[1] COPE Council. "COPE Guidelines: Retraction Guidelines". In: *Committee on Publication Ethics* (2019). DOI: `https://doi.org/10.24318/cope.2019.1.4`.

[2] Ridha Joober et al. "Publication bias: What are the challenges and can they be overcome?" In: *Journal of Psychiatry & Neuroscience* 37.3 (May 2012), pp. 149–152. ISSN: 1180-4882. DOI: `10.1503/jpn.120065`. URL: `http://dx.doi.org/10.1503/jpn.120065`.

[3] Kirils Makarovs and Peter Achterberg. "Contextualizing educational differences in vaccination uptake: A thirty nation survey". In: *Social Science & Medicine* 188 (2017), pp. 1–10. ISSN: 0277-9536. DOI: `https://doi.org/10.1016/j.socscimed.2017.06.039`. URL: `https://www.sciencedirect.com/science/article/pii/S027795361730415X`.

[4] Dmitry Malkov, Ohid Yaqub, and Josh Siepel. "The spread of retracted research into policy literature". In: *Quantitative Science Studies* 4.1 (Mar. 2023), pp. 68–90. ISSN: 2641-3337. DOI: `10.1162/qss_a_00243`. URL: `https://doi.org/10.1162/qss%5C_a%5C_00243`.

[5] Chephra McKee and Kristin Bohannon. "Exploring the Reasons Behind Parental Refusal of Vaccines". In: *The Journal of Pediatric Pharmacology and Therapeutics* 21.2 (Apr. 2016), pp. 104–109. ISSN: 1551-6776. DOI: `10.5863/1551-6776-21.2.104`. URL: `http://dx.doi.org/10.5863/1551-6776-21.2.104`.

[6] M. Lelinneth B. Novilla et al. "Why Parents Say No to Having Their Children Vaccinated against Measles: A Systematic Review of the Social Determinants of Parental Perceptions on MMR Vaccine Hesitancy". In:

Vaccines 11.5 (2023). ISSN: 2076-393X. DOI: 10.3390/vaccines11050926. URL: https://www.mdpi.com/2076-393X/11/5/926.

[7] Saad B. Omer. “The discredited doctor hailed by the anti-vaccine movement”. In: *Nature* 586.7831 (Oct. 2020), pp. 668–669. ISSN: 1476-4687. DOI: 10.1038/d41586-020-02989-9. URL: http://dx.doi.org/10.1038/d41586-020-02989-9.

[8] Hannah A. Roberts et al. “To vax or not to vax: Predictors of anti-vax attitudes and COVID-19 vaccine hesitancy prior to widespread vaccine availability”. In: *PLOS ONE* 17.2 (Feb. 2022). Ed. by Anat Gesser-Edelsburg, e0264019. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0264019. URL: http://dx.doi.org/10.1371/journal.pone.0264019.

[9] R. G. Steen. “Retractions in the scientific literature: is the incidence of research fraud increasing?” In: *Journal of Medical Ethics* 37.4 (Dec. 2010), pp. 249–253. ISSN: 0306-6800. DOI: 10.1136/jme.2010.040923. URL: http://dx.doi.org/10.1136/jme.2010.040923.

[10] The Center for Scientific Integrity. *The Retraction Watch Database [Internet]*. ISSN: 2692-465X. 2018. URL: http://retractiondatabase.org/.

[11] Richard Van Noorden. “More than 10, 000 research papers were retracted in 2023 — a new record”. In: *Nature* 624.7992 (Dec. 2023), pp. 479–481. ISSN: 1476-4687. DOI: 10.1038/d41586-023-03974-8. URL: http://dx.doi.org/10.1038/d41586-023-03974-8.

[12] AJ Wakefield et al. “RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children”. In: *The Lancet* 351.9103 (1998), pp. 637–641. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(97)11096-0. URL: https://www.sciencedirect.com/science/article/pii/S0140673697110960.