COMMENTARY

How many scientific papers are not original?

Michael Lesk[1]

Department of Library and Information Science, Rutgers University, New Brunswick, NJ 08901

Is plagiarism afflicting science? In PNAS, Citron and Ginsparg (1) count the number of authors who are submitting articles containing text already appearing elsewhere. They report disturbing numbers of authors resorting to copying, particularly in some countries where 15% of submissions are detected as containing duplicated material. I am on the editorial board of an Institute of Electrical and Electronic Engineers (IEEE) magazine, which also finds it useful to run all of the submissions through a plagiarism filter. What can be done about this?

In 1830, Charles Babbage deplored unreliable science. He discussed hoaxes, forgeries, data trimming, and "cooking" (selecting data to match a theory) (2). Today, doubtful papers may be plagiarized, invented, or mistaken. This paper documents problems at one extreme: straightforward plagiarism within one publisher. More complex deceptions can be found at the site retractionwatch.com, which includes, among other examples, invented or fraudulent data. Mistaken research was highlighted in an important study by Begley and Ellis, who found that it was impossible to replicate 47 of 53 oncology studies that they attempted to repeat (3). At a time when important scientific questions are under attack, we need to improve confidence in our publications.

How can we increase our level of trust in the scientific literature? In 2012, more than 2 million papers were published (4). They appear in publications ranging from highly competitive and prestigious journals such as *Nature, Science, Lancet*, and this journal, down to the predatory publishers listed in scholarlyoa.com who will print pretty much anything for a fee. University faculty, in particular, are encouraged to publish because the reward systems often depend on publication and citation counts as ways of evaluating merit. The h-index is the modern equivalent of the old saying "Deans can't read, they can only count." In some countries, having a paper accepted in a top journal can mean a cash bonus, with Zhejiang University offering a $30,000 payment to an author who publishes in *Science* or *Nature* (5, 6).

Given the incentives, it is hardly surprising that some authors are attempting to exploit the system. This can be surprisingly easy. Delgado et al. (7) explain how they created a half-dozen fake papers, with several hundred citations. One of the authors saw his citation count go up by a factor of 2 and his h-index increased from 10 to 15. Fans of bicycle racing may smile on reading that the fake papers were attributed to Alberto Pantini-Contador.

Refereeing, at least for some journals, is pretty shaky. As cited by Citron and Ginsparg, Bohannon (8) submitted a fake article to more than 300 open access journals, and more than half accepted it. Following up, he found that one of these journals had plagiarized its own description from a reputable journal in the same subject area. The scholarlyoa.com site attempts to catalog the doubtful publishers and their journals.

One bright spot in the Citron and Ginsparg paper is that plagiarism is concentrated: they note that a small number of authors produce a disproportionate share of the doubtful submissions.

Much more common than completely fake papers is the boosting of publication count by dividing one's reports into multiple short papers, an idea that has been called the "least publishable unit" since the 1970s. Some publishers or conference organizers join in the manipulation. Whilhite and Fong describe an editor who asked prospective authors to add citations to his journal to their articles to increase the impact factor of the journal (9).

Consequences

Deception and mistake can have real consequences outside of science. For decades, the UK educational system emphasized the "11-plus" examination, justified by a belief in the inheritability of intelligence that came from measurements by Sir Cyril Burt. Burt had studied what seemed to be a remarkable number of identical twins raised apart. His data were challenged soon after his death as too good to be true; the original notes were gone, and his coworkers could not be found. Although there has been argument back and forth, even his supporters have been defending him by saying he was careless rather than fraudulent and that other people studying genetics and intelligence have found about the same level of correlation (10).

More recently, two economists, Carmen Reinhart and Kenneth Rogoff, published a claim that economic growth slowed in countries whose national debt exceeded 90% of gross domestic product. After 2 y, they gave their spreadsheet to researchers at the University of Massachusetts, who found several errors; for example, the first few countries in alphabetical order had been left out of the calculation. A corrected spreadsheet did not show the same abrupt slowdown in growth, but the original paper had already been used to justify a change to budget-balancing policies in major economies (11).

Returning to the simpler problem of plagiarism, it can extend beyond individual papers. In 2009, a conference in Hainan, China, called itself the "International Joint Conference on Artificial Intelligence." That name is very familiar to artificial intelligence researchers as the title of a major conference held regularly since 1969. However, the conference with the long history met in Pasadena in 2009; the Hainan conference just borrowed the name. Perhaps it is not surprising that the Hainan conference included several papers that had come from the SCIGen chatterbot or some similar program. Here is a sentence from one abstract (since removed from IEEE Xplore): "Furthermore, it explored a pervasive tool for enabling pasteurization, which is used to show that context-free grammar and B-trees are largely compatible." Chatterbot output can now be detected automatically (12) and publishers find themselves, regrettably, forced to

Author contributions: M.L. wrote the paper.

The author declares no conflict of interest.

See companion article on page 25.

[1]Email: lesk@acm.org.

www.pnas.org/cgi/doi/10.1073/pnas.1422282112

use such software, as well as anti-copying utilities.

Plagarism would matter less if counting articles was less significant than understanding them. ArXiv at least does not claim to referee submissions; anyone using it knows that they have to read and evaluate the content for themselves. This, of course, transfers the burden of judgment from a small number of referees to the much larger number of potential readers. In addition, many of those readers may be students, or in a different discipline, and be less able to evaluate a paper. This is why we have the current publication system, but it is being abused by researchers who know that for some purposes, the main question being asked of a candidate for hiring or promotion is "how many articles?"

Mere number of publications is not what is really important. When challenged as a "half-wit," the Roman emperor Claudius, at least in the British Broadcasting Corporation version of his life, replied that it is quality rather than quantity of wits that matters (13). Similarly, the National Science Foundation asks those who submit proposals to list five important and relevant papers and not to attempt to drown the referees in dozens (or hundreds) of articles.

Fortunately, one bright spot in the Citron and Ginsparg paper is that plagiarism is concentrated: they note that a small number of authors produce a disproportionate share of the doubtful submissions. In addition, those articles are not the heavily cited ones, suggesting that they have less influence. Also, there are many important countries where the plagiarism rate is low. Conversely, the methodology of the paper relies on exact text overlap; it will not detect, for example, an article translated from another language, nor one which paraphrases but adds nothing to its source.

Possible Actions

What can we do? This paper observes a strong cultural connection with plagiarism: there are some countries in which 15% of the submissions to arXiv are plagiarized, and others in which very few papers are copying from others. Can the scientific community, with some combination of carrots and sticks, encourage the institutions in all countries to enforce standards? There are very few individual scientists today, and approaching the institutions might be the best way to affect a change in attitude.

For example, recently I received a request from someone in Asia who wanted to be a postdoctoral researcher in our department in the United States. I took the first two paragraphs of his research statement and found them on a commercial website of a US company. Should I have told this to the head of his institution? Right now, we don't do that, partly out of politeness and partly out of fear of lawsuits. However, when Citron and Ginsparg write that some of the people whose plagiarism is detected reply by asking to be told which parts were found to be copied, presumably to learn how to evade detection in the future, one despairs.

For experimental studies, the move to requiring data availability will be a step forward. If an author did not actually write the paper under discussion, presumably that author does not have the data behind it. The data can be copied as well, but that offers another chance for automated tools to spot the duplication, and one where paraphrasing is more complicated.

ArXiv is trying to motivate authors by flagging papers that contain overlap. Readers are then on notice that the paper has a problem; unfortunately, authors do not necessarily react with shame or withdrawal. Some ignore the flag, and some say that what they are doing is acceptable practice. These responses suggest that some additional response is needed (although Citron and Ginsparg do not say how many authors respond to the warning in which way).

Nature published a discussion on plagiarism 2 y ago, and in it, Zhang and McIntosh suggested keeping a blacklist of individuals (14). They note that this should be a multipublisher effort and that it is unclear who would run it or pay for it (14). I would suggest one further step: identify departments, and perhaps institutions, where the problems are arising. Publishers should suggest that they will blacklist the entire department (or, if need be, the institution). Intermediate forms of punishment are possible, such as delaying publication rather than denying it entirely.

In summary, this paper describes the scope of plagiarism within arXiv. The good news is that the tools used to detect plagiarism work effectively and efficiently, the copied papers are concentrated by author and by country, and the copied papers are less cited. The bad news is that the problem is real and in some countries severe. ArXiv is now identifying the papers that have substantial overlap and is waiting to see if that affects the submissions. Perhaps the publishing community as a whole should be preparing to see if stronger steps are needed.

1 Citron DT, Ginsparg P (2015) Patterns of text reuse in a scientific corpus. *Proc Natl Acad Sci USA* 112: 25–30.

2 Babbage C (1830) *Reflections on the Decline of Science in England, and on Some of Its Causes* (B. Fellowes, London).

3 Begley CG, Ellis LM (2012) Drug development: Raise standards for preclinical cancer research. *Nature* 483(7391):531–533.

4 Reich ES (2013) Science publishing: The golden club. *Nature* 502(7471):291–293.

5 Davis P (2011) Paying for impact: Does the Chinese model make sense? Available at: scholarlykitchen.sspnet.org/2011/04/07/paying-for-impact-does-the-chinese-model-make-sense/. Accessed November 24, 2014.

6 Shao J, Shen H (2011) The outflow of scientific papers from China: Why is it happening and can it be stemmed? *Learn Publ* 24(2):95–97.

7 Lopez-Cozar E, Robinson-Garcia N, Torres-Solinas D (2013) Manipulating Google Scholar citations and Google Scholar metrics: Simple, easy and tempting. Available at: arxiv.org/abs/1212.0638. Accessed November 27, 2014.

8 Bohannon J (2013) Who's afraid of peer review? *Science* 342(6154):60–65.

9 Wilhite AW, Fong EA (2012) Scientific publications. Coercive citation in academic publishing. *Science* 335(6068):542–543.

10 Plucker JA, Esping A, eds. (2014) The Cyril Burt affair. *Human Intelligence: Historical Influences, Current Controversies, Teaching Resources*. Available at: www.intelltheory.com. Accessed November 23, 2014.

11 Krugman P (2013) How the case for austerity has crumbled. *The New York Review of Books*. Available at: www.nybooks.com/articles/archives/2013/jun/06/how-case-austerity-has-crumbled. Accessed on November 27, 2014.

12 Labbé C, Labbé D (2013) Duplicate and fake publications in the scientific literature: How many SCIgen papers in computer science? *Scientometrics* 94(1):379–396.

13 Pullman J (1976) *I, Claudius* [television production], director Wise H (British Broadcasting Corporation).

14 Zhang Y, McIntosh I (2012) How to stop plagiarism: Blacklist repeat offenders. *Nature* 481(7379):22.