POLICING PLAGIARISM

The internet has made both copying other people's work and detecting plagiarism much easier. **Michael Cross** looks at some of the tools used to tackle plagiarism

In the internet age, copying someone else's work can be as simple as clicking and dragging a computer mouse over a few plausible paragraphs. By the same token, the world wide web makes fraud easy to detect. Over the past decade, a range of software products has become available for detecting plagiarism, especially by students. However, experts are questioning whether Britain's strategy for detecting academic fraud is the right one for catching the most damaging types of misconduct.

There is no evidence that plagiarism is becoming more prevalent in research. But there is no doubt that plagiarism happens, perhaps because of mindsets acquired in education.[1] The Committee on Publication Ethics, an international forum for editors of peer reviewed journals, has discussed "30 or 40" alleged cases of research plagiarism over the past 10 years, says its chairman, Harvey Marcovitch.

The most common type of plagiarism is where a relatively junior researcher copies passages from published work into a paper. Such authors may claim they did not know they were doing anything wrong, especially if English is not their first language and they were educated in a more hierarchical culture. "Often the explanation is 'My English is very poor, so I thought it was better to use the words of someone senior to me'," says Marcovitch.

There are, however, more serious cases. Marcovitch recalls a case when, on one occasion, as editor of *Archives of Disease in Childhood*, he was contacted by a reader pointing out that a paper had already appeared in an East African journal. "The only difference was that the order of the authors had been changed. The 'author' had clearly swiped the entire paper."

Defining plagiarism

The Council of Science Editors, which promotes ethical practices in science publishing, defines plagiarism as "a form of piracy that involves the use of text or other items (figures, images, tables) without permission or acknowledgment of the source of these materials." Though plagiarism usually involves the use of materials belonging to others, the term can apply to duplicate publication—researchers duplicating their own previous reports without acknowledging that they are doing so. Marcovitch says such "self plagiarism" can be particularly pernicious: "Cases of duplicate publishing often involve people who are quite senior. Sometimes they're offended at the suggestion that they've done something wrong." But the consequences can be serious. "If it's original research that gets recycled, it may subvert the scientific process by suggesting an evidence base different from what is the case."

Although such deception might be picked up by peer reviewers, internet technology can alert even a non-specialist editor that something may be wrong by matching a paper against previously published work.

This type of check can be done in three ways. One is simply to paste passages from the suspect work into a search engine and look for exact matches. A less laborious method is to use a commercial service, which compares the work with a more comprehensive archive than that held by general search engines, and graphically highlights passages bearing similarities to other works. This can pick up cases where the plagiarists have tried to cover their tracks by replacing words with synonyms. Critics say it's dangerous to rely on a single detection tool, especially from a commercial source. They favour the use of freely available non-commercial programs, especially those incorporating "fuzzy" techniques to find non-identical matches for further investigation. But no computer program is on its own capable of deciding whether or not an author is guilty of plagiarism.

Relying on a single tool

The UK academic community's approach to battling plagiarism relies heavily on a single commercial program, Turnitin, developed and marketed by a US supplier, iParadigms. The company claims its database contains over 40 million student papers. The Joint Information Systems Committee (JISC), which supports education and research in IT in higher education, picked the system in 2002 when it offered free access to the software through its plagiarism advisory service, to which 80% of higher education institutes subscribe.

Free access ended in September this year, when JISC split the service in to two, JISC-iPAS to focus on technology and the Academic Integrity Service to look at other concerns.[2] Access to Turnitin and a sister product, iThenticate, aimed at commercial publishers, are now available through a commercial spin-off, Northumbria Learning (www.northumbrialearning.co.uk). The com-

"If it's original research that gets recycled, it may subvert the scientific process by suggesting an evidence base different from what is the case."

pany does not publish prices but says that a typical licence would cost "around 2-3k, plus a charge per page depending on the volume of submissions."

Fiona Duggan, head of advice and guidance at the Academic Integrity Service, says that independent studies have confirmed Turnitin's position as the most effective single tool. A study by the National Computing Centre ranked the product first among 11 contenders. But she says that the system is not foolproof and the academic community should be wary of using a single technique to catch plagiarism, because suspect practices that fall through the net may become acceptable behaviour.

Fintan Culwin, professor of software engineering education at London South Bank University, shares this reservation. "Unfortunately many people think [Turnitin is] the tool. If the only tool you have is a hammer, everything starts to look like a nail." Culwin warns that an increasing reliance on commercial tools is taking academic integrity "beyond its comfort zone," and reliance on a single supplier is especially risky.[3]

One worry is that commercial suppliers treat their accumulated corpus of work as a proprietary asset, which could hinder open research. Another is sustainability: in the US, Turnitin has faced claims that it violates copyright by retaining copies of student papers in its archive.[4]

Alternative tools

As an alternative, Culwin suggests that institutions look to free detection tools, including those of his institution's own Centre for Interactive Systems Engineering. They automate the first two stages of the four stage process of detecting plagiarism: collection and analysis (detection of non-originality). The final two stages, confirmation and investigation, require human intelligence, he says.

One tool is VAST—the visualisation and analysis of similarity tool—which presents similarities between documents in a more visual way than previous products. By overlaying digital representations of two documents, it shows the extent of similarity as a dark smudge running diagonally down the display.

Another development is FreeStyler, a stylometric program that detects uncharacteristic shifts in writing style within a document, possibly indicating that passages have been lifted from elsewhere. Stylometric programs assess factors such as a work's reading age, spelling conventions, and punctuation and look for passages where all these change simultaneously. This is a good indication that a passage has been copied, but does not produce the same quality of evidence as showing the source material, unless a match can be found on the web.

Whatever software is used, the decision on whether to accuse a student of cheating must be a human one, and handled with care, says Culwin. "It can be very unsettling for students, especially as I have several from the sort of families where letters might be opened by their parents. We can't be hanging judges."

Culwin says that students are now learning that blatant plagiarism will be caught by software checks. "The amount of non-original material is decreasing, and the way it is used is qualitatively changed—students are now much more likely to attempt to reference plagiarised material."

Publishing houses are also turning to technology to police plagiarism. One scheme currently being piloted is CrossCheck, a collaboration between iThenticate, a tool that enables publishers to check the originality of documents and manuscripts, and CrossRef, a citation linking system that allows a researcher to click on a reference citation on one publisher's platform and link directly to the cited content on another publisher's platform.

One problem for publishers checking if manuscripts have been copied is that the source material they may have plagiarised may be behind access controls. CrossCheck allows publishers to work together to allow mutual access to their content in order to detect plagiarism. With the publisher's permission, any article that has a unique digital identifier on CrossRef will be indexed and held on the iThenticate's system. Publishers are then able to check for replication both before and after publication. Prepublication checks occur after a manuscript has been submitted and a check is run to see the percentage overlap with other publications. Postpublication checks enable publishers to find out if their content has been lifted and used elsewhere.

Bespoke work

The bad news is that, rather than risk being caught copying, cheating students are now likely to turn in original work written by paid confederates. Simple plagiarism detection tools will not identify this kind of cheating, as the bespoke work matches no original document in the corpus. Detecting this kind of paid impersonation is one reason for Culwin's interest in stylometry. But detecting fraud is not as simple as showing changes in a student's writing style—something that tends to happen during a course of education, ideally towards more straightforward language—but will need new systems to identify unknown authors from the style of their work, perhaps with techniques similar to those used by scholars investigating the authorship of historic documents..

Whatever progress is made in software development, everyone agrees that the day of fully automated detection of publishing misconduct is far off.

Duggan says that the most successful applications of software have been within broader institutional changes. "Software is just a tool. It has to be part of a much bigger change within the culture. For us a successful implementation is an institution which has looked at its policies and revised them and was able to develop a culture where students are aware of appropriate behaviour."

Marcovitch says that one answer may lie in implanting the norms of ethical behaviour much sooner than at present. "The problem is that the real crooks are not going to be put off by ethics guidelines."

Michael Cross is a freelance journalist in London
michaelcross@fastmail.fm

Competing interests: None declared
Provenance and peer review: Commissioned; not peer reviewed.

1 Gerhardt D. 2006. Plagiarism in cyberspace: learning the rules of recycling content with a view towards nurturing academic trust in an electronic world. *Richmond J Law Technol* 2006:12(3):10. http://law.richmond.edu/jolt/v12i3/article10.pdf
2 Scaife B. *IT consultancy plagiarism detection software report for JISC Plagiarism Advisory Service*. 26 September 2007. www.jiscpas.ac.uk/documents/resources/PDReview-Reportv1_5.pdf
3 Lancaster T, Culwin F. 2004. Using freely available tools to produce a partially automated plagiarism detection service. In: Atkinson R, McBeath C, Jonas-Dwyer D, Phillips R (eds). *Beyond the comfort zone: proceedings of the 21st ASCILITE Conference*. Perth, 2004:520-9. www.ascilite.org.au/conferences/perth04/procs/lancaster.html
4 Glod M. McLean students sue anti-cheating service. *Washington Post* 2007 March 29:B05. www.washingtonpost.com/wp-dyn/content/article/2007/03/28/AR2007032802038.html.