Using digital forensic techniques to identify contract cheating: A case study

Clare JOHNSON, Ross DAVIES
University of South Wales, United Kingdom

Academics typically use two methods for detecting plagiarism: a tool such as Turnitin®, which provides a suite of online educative and evaluation tools including a section that checks for originality of work submitted (www.turnitin.com), or their knowledge of the student and likely standard of work as a flag for what to expect – an outstanding piece of written work from a student that struggles to write a bullet point on a post it note is likely to raise the attention of the assessor. Other techniques include the use of online search tools, where unusually phrased sentences in an assignment, which may seem out of character for the student or within the context of the rest of the assignment, can be pasted into Google to see if a match can be found.

In their paper of 2009, Bretag and Mahmud conclude that electronic detection provides an effective starting point in detecting plagiarism but that this must be "combined with considerable manual analysis and subjective judgement". Identifying contract cheating introduces further problems: the work may be original and of good standard – it just isn't written by the person who has submitted it. "Educators and researchers working in the field of academic integrity agree that electronic detection is not the solution to eliminating plagiarism" (Bretag & Mahmud, 2009), whilst Rogerson (2017) suggests that "Some knowledge of the practices of students ... can be useful to identify instances of potential contract cheating". This can be difficult in large classes or where assessors do not know the students they are assessing.

Indeed, Turnitin recognises that whilst their detection tools are hugely beneficial, they are still limited in their ability to detect contract cheating. They are currently developing 'Authorship Investigation', which will use stylometry and other semantics to help establish authorship of a document.

The researchers in this project both work in the academic Cyber Security department of a UK Higher Education Institution. They have a particular interest in teaching and learning, and both lecture on digital forensics, teaching students how to carry out digital forensic investigations to a level whereby they could feasibly present an expert witness statement in court. Topics include the use of digital forensic tools such as Autopsy (free) and FTK (proprietary). Steganography techniques are also taught.

Contract Cheating Case Study

The researchers were alerted to an alleged instance of contract cheating by the contracted author (hereafter referred to as Ms A). Ms A emailed the department saying that one of the students of the University had used a contracting website to request some work to be done and noted that the person in question has 'a habit of not paying after collecting the scripts' (personal communication, 21 January 2018). Having failed to receive payment, Ms A investigated the assignment brief in more detail and was able to determine which University the assignment came from and the contact details for the department. She provided screenshots of the contract being negotiated, and the work that she produced in response and sent these to the department.

On receipt of these documents a quick comparison was carried out with the student submission, which showed that there were significant similarities between the work of the contractor and the student. Following standard academic process for the University, the student was referred to an Academic Misconduct hearing where he confessed that he had posted the brief on a contracting website and presented the work produced as his own. The reliability of the allegation against the student is therefore not in question. Publication of the findings of this research has been approved by the Faculty's ethics champion.

Digital Forensics Techniques in other situations

During the literature review it was possible to locate various articles that discuss forensic techniques similar to those used in this case study, but for very different purposes, such as Fu, Sun, Liu & Li (2011) for checking originality of a document in relation to copyright issues and research by Xiang, Sun, Liao, & Wang (2016), who discuss the use of these techniques for hiding data within a Word document (steganography). The methods described below can be used in criminal investigations, but no evidence was found to suggest that they are ever used in establishing that contract cheating has occurred.

Techniques used

There are some very simple tools which can be used to help establish ownership of a document created in Microsoft Word. In Word 2016, Document Properties can provide some basic information such as file size, number of pages, total editing time, company (if used), author and last modified by. As long as the document is still in Word format (and not PDF), these can be easily viewed by opening the file normally and selecting File, Info and Properties.

In order to investigate more thoroughly, an understanding of how a Word document is built is required. A Word document is essentially a collection of other files, gathered together and compressed into a single 'docx' file – much like a zip file which contains a number of documents compressed for sending over the Internet. In most cases, it would never be necessary to decompress a 'docx' file. However, these files, when decompressed, reveal some very useful information about the origins of the work. They contain metadata, document properties, formatting, hyperlinks, and the text itself. Most of these are not of interest to us at this stage. This research focuses on the document.xml file, which in this case reveals some interesting features.

Discussion

Word documents are designed with author collaboration in mind and have a facility to detect specific edits to the contents (e.g. text and images). These edits are marked with values called "Revision Save Identifiers", more commonly referred to as rsid. These values are randomly generated but increment throughout a document's life span, for example when a revision is made, or when the document is saved. This allows two authors to work on the same document where changes are merged based on these values. This information proves valuable when reviewing a document submitted by a student suspected of contract cheating, and having developed a simple tool for analysis the researchers were able to review the rsid tags in the case study submission.

When a student writes an assignment they will typically go through a series of activities: research, brainstorming, developing content, editing, adding citations and figures, proof reading and corrections. On reviewing the document.xml file of a genuine assignment submission, it is clear to see all the edits that take place during this process. Edits are represented by rsid values wrapped around the text that has been edited and clearly show where someone has added or amended content over a period of time.

Conversely, when a student contract cheats, they will receive a completed assignment written by the contractor. It is unlikely that they would submit this document in its original form, as the metadata would indicate that the author is not the student (and for cyber students, this would be common knowledge). It is more likely that paragraphs will be imported from the contractor's work into a new document created by the student. At the point of pasting, rsid values are stripped out automatically, leaving one rsid edit tag for a whole paragraph. This appears highly unusually for an original piece of work. A student will then carry out some further edits: adding their name, university details, changing the formatting, removing or amending work they are not entirely happy with and adding to the content. Again, these edits or word substitutions are very clear.

Through this is it possible to see on the contracted work that large chunks of text 'appear' with only minor edits of single words / phrases, all completed on a single edit. This is in contrast to an original submission, which is littered with edits throughout, with almost no large runs of text. Further analysis makes it possible to determine the order of edits and this is an area that will be further researched.

Summary

Whilst there are limitations with the above analysis – in particular that only one contracted submission has been fully reviewed and compared with a number of original submissions, initial findings suggest that further analysis would yield very interesting results and add to the evidence that contract cheating has occurred. If this can be formalised and turned into a practical tool, it could be used to support academic staff in identifying cases of contract cheating much more easily.

Keywords: contract cheating, digital forensics, xml, plagiarism detection.

References

Bretag, T., & Mahmud, S. (2009). A Model for Determining Student Plagiarism: Electronic Detection and Academic Judgement. *Journal of University Teaching and Learning Practice*, 49-60.

Fu, Z., Sun, X., Liu, Y., & Li, B. (2011). Forensic investigation of OOXML format documents. *digital investigation*, 8(1), 44-55. doi:10.1016/j.diin.2011.04.001

Rogerson, A. M. (2017). Detecting contract cheating in essay and report submissions: process, patterns, clues and conversations. *International Journal for Educational Integrity*.

doi:10.1007/s40979-017-0021-6

Xiang, L., Sun, C., Liao, N., & Wang, W. (2016). A Characteristic-Preserving Steganographic Method based on Revision Identifiers. *International Journal of Multimedia and Ubiquitous Engineering*, 11(9), 29-38.