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

Clare Johnson

Dr Ross Davies

University of South Wales, Treforest Campus, Pontypridd, CF37 1DL

Corresponding author: Clare Johnson, clare.johnson@southwales.ac.uk, 01443 483246

ORCID ID: Clare Johnson 0000-0001-5869-957X

ORCID ID: Dr Ross Davies 0000-0002-5460-9837

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

Abstract

Contract cheating is a major problem in Higher Education because it is very difficult to detect using traditional plagiarism detection tools. Digital forensics techniques are already used in law to determine ownership of documents, and also in criminal cases, where it is not uncommon to hide information and images within an ordinary looking document using steganography techniques. These digital forensic techniques were used to investigate a known case of contract cheating where the contracted author has notified the university and the student subsequently confirmed that they had contracted the work out. Microsoft Word document use a format known as Office Open XML Format, and as such, it is possible to review the editing process of a document. A student submission known to have been contracted out was analysed using the revision identifiers within the document, and a tool was developed to review these identifies. Using visualisation techniques it is possible to see a pattern of editing that is inconsistent with the pattern seen in an authentic document.

Keywords: Contract cheating, plagiarism, forensics, detection, academic integrity, OOXML

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

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. In addition, one of the authors has a special interest in plagiarism and contract cheating.

Contract Cheating Case Study

A known case of contract cheating was used for this case study. The case was not in question – emails from the contracted author (hereafter referred to as the Contract Author were available for review, showing that the Contract Author had contacted a university department claiming that a student from that 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, the Contract Author investigated the assignment brief in more detail and was able to determine which university the assignment came from, and from there the contact details for the relevant department. The

Contract Author provided screenshots of the contract being negotiated, and the work that they had produced in response and sent these to the relevant university department.

The student submission was also available for review. A quick comparison between the contracted work and the student submission was carried out, which showed that there were significant similarities between the work of the Contract Author and the student. Following standard academic process for the university in question, the student was referred to an Academic Misconduct hearing where the student confessed that they had posted the brief on a contracting website and presented the work produced as their own. The reliability of the allegation against the student is therefore not in question. Ethical approval for the discussion surrounding this case has been granted.

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. (Xiang, Sun, Liao, & Wang, 2016) suggest the use of extensible mark up language as a cover medium to 'transmit secret information by offenders' and discuss tools and techniques to detect deliberately hidden information. Similarly, Castiglione, D'Alessio, De Santis & Palmieri (2011) explain how data can be hidden in a single file using the OOXML format. Jeong & Lee (2017) discuss the use of digital forensics techniques in relation to version history to suggest theft of intellectual property though establishing document contents creation, modification, deletion and copy.

Establishing Ownership of a Word Document

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.

Structure of a Word Document

In order to investigate more thoroughly, an understanding of how a Word document is built is required. Microsoft Word uses the 'Office Open XML Format' (OOXML) format. 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 meta data, document properties, formatting, hyperlinks, and the text itself.

This research focuses on the document.xml file, which in the case of contract cheating 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). In the document.xml file, these edits are marked with values called "Revision Save Identifiers", more commonly referred to as rsid. This allows two authors to work on the same document where changes are merged based on these values. 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 (ISO, 2016). 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: brainstorming, research, 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 codes precede 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 document properties would indicate that the author is not the student, and this information is readily available to review. 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 (see Digital Forensic Analysis section below). 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 and again, these edits or word substitutions are very clear. Runs of edits appear marked with an rsidR tag, where a series of small edits are made in one editing session (perhaps before an autosave, or a user invoked save).

Through this analysis 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 run. This is in contrast to an original submission, which is littered with edits throughout, with almost no large runs of text. The hypothesis for this research is therefore that it is possible to detect contract cheating through using digital forensics techniques which establish an unorthodox pattern of editing consistent with contracted out work and inconsistent with original work.

Practical example of Revision Identifiers

To test this theory, a new blank document is opened in Microsoft Word 2016 v 16.0.4738.1000. The text 'Digital Forensics.' is entered. The file is saved with the filename 'Digital

Forensics.docx' and closed. The Word document is viewed in File Explorer and the contents decompressed using a zip tool. The attribute values in the document.xml file were reviewed:

```
<w:p w:rsidRDefault="0009533F" w:rsidR="00CE3196">
  - <w:r>
      <w:t>Digital Forensics.</w:t>
    </w:r>
</w:p>
```

Fig. 1. Partial code from 'document.xml' file following the document creation.

Attribute description:

Attribute Reference	*Description*	*Value*
rsidRDefault	Default Revision Run Identifier	0009533F
rsidR	Revision Identifier for Run	00CE3196

In the second example, the file was reopened and the word 'Forensics' changed to 'Evidence'. This modification results in the creation of an additional run:

```
<w:p w:rsidRDefault="0009533F" w:rsidR="00CE3196">
  - <w:r>
      <w:t xml:space="preserve">Digital </w:t>
    </w:r>
  - <w:r w:rsidR="00694EF2">
      <w:t>Evidence</w:t>
    </w:r>
  - <w:r>
      <w:t>.</w:t>
    </w:r>
    <w:bookmarkStart w:name="_GoBack" w:id="0"/>
    <w:bookmarkEnd w:id="0"/>
</w:p>
```

Fig. 2 Partial code from 'document.xml' file following an edit.

Attribute Reference	*Description*	*Value*
rsidRDefault	Default Revision Run Identifier	0009533F
rsidR	Revision Identifier for Run	00CE3196
rsidR	Revision Identifier for Run	00694EF2

The edit where the text change has occurred is identified by a new run value (rsidR) of 00694EF2. This code demonstrates that the rsidR value increments as edits are made – though edits made in a single run (i.e. before a save or autosave) will share the same rsidR value. The

original rsidR value is still visible because the word ‘Digital’ has not been altered during the second edit.

Digital Forensic Analysis

Using a tool created by one of the authors, the above document (‘Digital Forensics.docs’) was analysed. The tool, written in PhP, uses visualisation techniques to extract the rsid values and colour code them for ease of viewing. In the figure below, the first editing run is highlighted in red with the value 00CE3196, and the second editing run is highlighted in blue with the value 00694EF2. The tool in fact uses the hexadecimal value of the run to fix the colour for visual purposes – these have been changed in some cases to aid clarity for the reader.

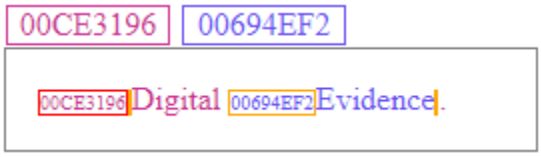

Fig. 3. Tool interpretation of the file.

A paragraph from this paper was also uploaded to the tool to provide an example of how a typical research paper might look, given usual editing processes. This is referred to as the Author Work Example Extract:

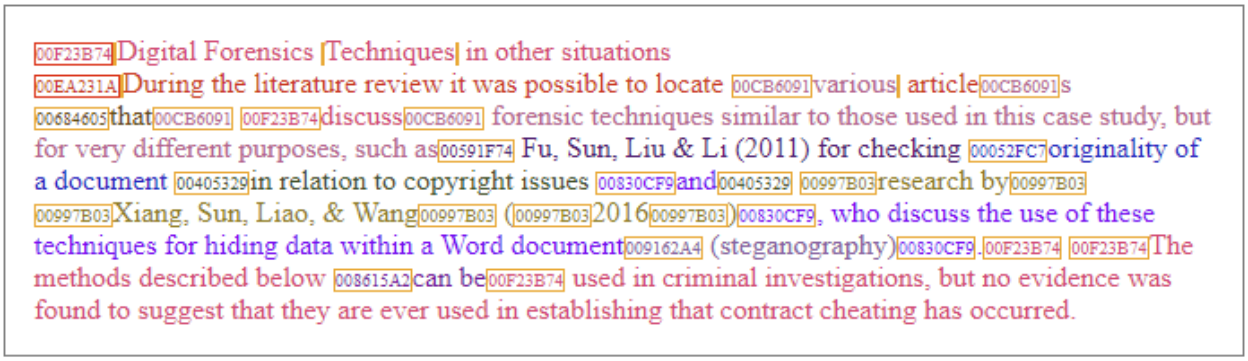

Fig. 4. Author Work Example Extract showing rsidR values.

Here it is possible to see a number of run values (rsidR) as the writing, referencing and editing process evolves. In this paragraph alone there are twelve unique rsidR values for a paragraph and title consisting of 106 words in total.

On the above paragraph there are 12 unique RSID values. This gives the following ratios:

Words	Unique rsidR values	Ratio of edits to words
106	12	11.32%

Fig. 5. Author Work Example Extract ratio of edits

A sort was carried out on the rsidR hexadecimal values to see whether an order of edits could be established, as suggested by the specification (ISO, 2016). However, a number of differing combination were tried (extracting pairs or octets of hex values and inverting them) but it was

not possible to confirm a timeline in this way. This may be an area worth reviewing for future research as it would be useful to be able to create a timeline of edits on a suspected contracted assignment.

Analysis of copied work

The sampled paragraph analysed above was sent from one of the authors of the paper to the other. On receipt, the second author saved the document, changed one word ('issues' changed to 'infringement'), and then resaved the work. This reflects the typical behavior of a contracting student and is referred to hereafter as Contracted Work Example Extract. This sample was uploaded to the forensic tool.

The results are shown below:

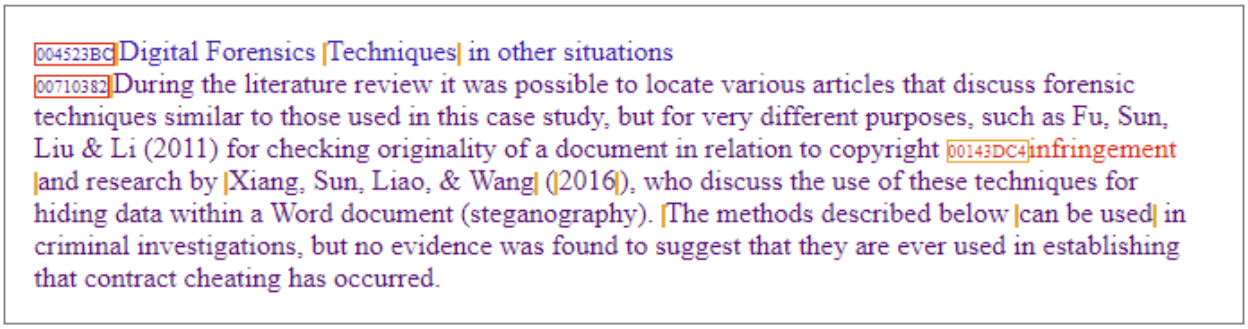

Fig. 6. Contracted Work Example Extract showing rsidR values.

On this paragraph there 3 unique rsidR values. This gives the following ratio, which is much lower than that of the original work.

Words	Unique rsidR values	Ratio of edits to words
106	3	2.83%

Fig. 7. Contracted Work Example Extract ratio of edits

Case study

In our case study, we compared the work written by the contracted author to the work submitted by the student. This yields the following results:

	Words	Unique rsidR values	Ratio of edits to words
Contract Author	3685	88	2.39%
Student submission	4172	15	0.36%

This doesn't provide very meaningful data. On closer inspection, the reason for this is that when a student uses contracted work, they typically change only one or two words per run in the author's document. This is to change connective words, words that have regional variations,

contextual variations, spelling errors etc. The contracting student doesn't tend to edit lengthy runs of text. The results above don't take this into account, and therefore the data appears not to suggest anything untoward.

What is more compelling is a visual representation of the document. In the example below, the same paragraph was visually represented using different colours to identify the rsidR runs. Using this technique yields the following images:

Original paragraph:

Fig. 8. Author Work Example Extract visualised

Contracted paragraph:

Fig. 9. Contracted Work Example Extract visualised

This is amplified in the case study example, where large blocks of colour appear with minor edits which are all completed in the same editing run. Inspection of the Contracted Author work shows that these edits are to amend spelling errors, remove quotes around words, to remove surplus spaces and similar.

Statistical analysis of the data gives the following results:

	Contracted Author	**Student Submission**
Word count	3685	4172
Unique rsidR values	88	15
Ratio of unique edits to words	2.39%	0.36%
Total number edits	193	55
Ratio of total edits to words	5.24%	1.32%
Ratio of rsidR to total edits	45.60%	27.27%

Fig. 10. Full document comparison of ratios

A larger section from the Contracted Author's work and the Student Submission are shown below. Whilst difficult to statistically verify this, the visual appearance highlights the difference between the originally edited work and the copied work. The author's used examples of their own original work as a comparison and saw a similar pattern of editing as the Contracted Author's, as opposed what is seen in the Student Submission.

Fig. 11. Contract Author's work visualised

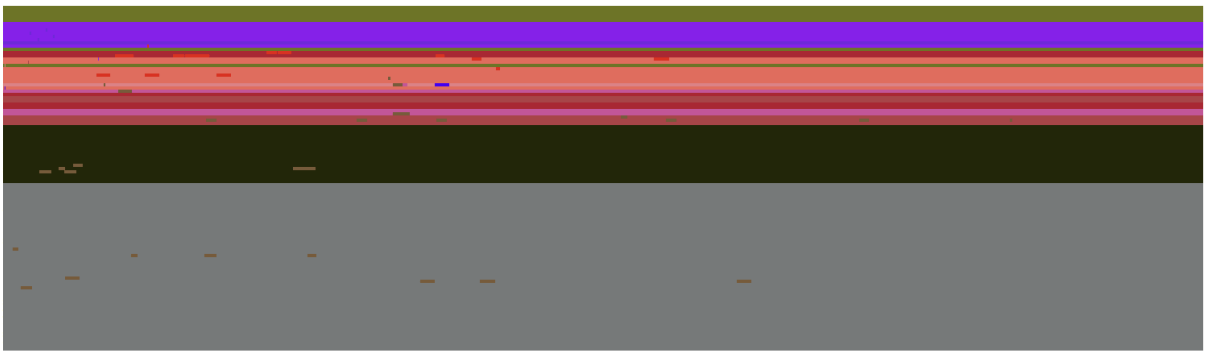

Fig. 12. Student Submission work visualised

Limitations

The main limitation of this case study is that at present, only one full document has been analysed. Whilst this is known to definitely be contracted work, further examples will be need to establish consistency in the findings.

In addition, it is important to note that there may be alternative reasons for a document appearing with this structure of edits (or lack thereof). For instance, a student may work on a document at length, and then copy and paste the entire work into a separate file – perhaps because it was originally embedded with some other work, or because it was transferred into another document. The analysis of the rsidR values also needs further development, as it currently fails to take into account a large number of minor edits across a large section of text – which would be consistent with minor edits to contracted work. It also falsely represents section headings and lists which show as a longer block of colour than they should because of the way are interpreted by the visualisation. This technique must therefore be used in conjunction with other more traditional methods of detection.

Finally, the authors of this paper were able to compare the Contracted Author's work with the Student Submission – this is not a luxury that would be available when inspecting submitted work, so it will be necessary to quantify this in some way in order to make the tool more reliable

Summary

In spite of the limitations indicated above, the finding from this research suggest that analysis using these techniques could add to the evidence that contract cheating has occurred. The visual representation gives a clear indication of the way the document has been edited, though the statistical analysis at this stage is in its infancy and doesn't fully corroborate what can be seen visually. If this method 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.

Bibliography

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.

Castiglione, A., D'Alessio, B., De Santis, A., & Palmieri, F. (2011). New Steganographic Techniques for the OOXML File Format. In A. M. Tjoa, G. Quirchmayr, I. You, & L. Xu (Eds.), *Availability, Reliability and Security for Business, Enterprise and Health Information Systems* (Vol. 6908). Springer, Berlin, Heidelberg. doi:https://doi.org/10.1007/978-3-642-23300-5_27

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

ISO. (2016). Information technology - Document description and processing languages - Office Open XML File Formats. *ISO/IEC 29500-1*, 1060-1063.

Jeong, D., & Lee, S. (2017). Study on the tracking revision history of MS Word files for forensic investigation. *Digital Investigation, 23*, 3-10. doi:10.1016/j.diin.2017.08.003

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.

Zhangjie, F., Xingming, S., & Jie, X. (2015, October). Digital Forensics of Microsoft Office 2007–2013. *Journal of Communications and Networks, 17*(5), 525-533. doi:10.1109/JCN.2015.000091