Understanding academic integrity and plagiarism in the digital age: Can digital forensics techniques help prevent and detect academic misconduct?

Clare Johnson

A submission presented in partial fulfilment of the requirements of the University of South Wales / Prifysgol De Cymru for the degree of Doctor of Philosophy

February 2023
Candidate declaration

This is to certify that, except where specific reference is made, the work described in this thesis is the result of my own research. Neither this thesis, nor any part of it, has been presented, or is currently submitted, in candidature for any other award at this or any other University.

Signed

Clare Johnson
Candidate

Date

09 February 2023
Acknowledgements

Thank you to my supervisors, all of whom were generous with their time, knowledge and expertise. These include Dr Mike Reddy, who provided moral guidance and whose knowledge of the field of academic integrity helped shape my understanding in a holistic way, and who was always available when I needed encouragement; Dr Ross Davies, whose ability to read and review my programming code was astounding; Dr Susan Haywood for her wise guidance and attention to detail, and Dr Andrew Ware for helping to guide me over the finish line. All these I consider friends as well as colleagues and supervisors.

I would like to thank the students that took the time to complete my surveys, and the staff at the University of South Wales who were willing to allow me to disrupt their classes to do this, as well as the student casework team at the University who assisted me in accessing student files for analysis.

Finally, I would like to thank my family, my husband Mike, and my children, Lauren and Ffion and Zac, all of whom have probably heard more about my PhD than they really cared to, but who supported me every step of the way nonetheless. Thank you.
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Abstract

Academic integrity is integral to the education system. Society relies upon the integrity of qualifications awarded to individuals during their studies as evidence of their knowledge and ability. Sadly, academic integrity is not always upheld, and for multifarious reasons, students find themselves engaging in academic misconduct, whether that is by copying exam answers from a peer, collaborating on an assignment that was meant to be individual work, copy-and-paste plagiarism, or for some, when the challenges of completing work on time become too great, turning to an essay mill to complete their work on their behalf. Approaches to ensuring academic integrity focus on prevention, detection and penalty, and these can be addressed through culture, strategy and policy. This thesis takes a holistic approach to academic integrity, by firstly exploring students’ perceptions of academic integrity and academic misconduct, and then secondly by developing a novel approach to the detection of plagiarism and contract cheating. It is this latter theme that forms the substantive part of the research, through the development of a novel detection tool, ‘Clarify’, that uses a cyber security related digital forensics approach to the detection of plagiarism, contract cheating, collusion and the much more recent problem of artificial intelligence generated essays. Existing plagiarism detection tools focus on the textual content of the document, searching databases and repositories for text matches, and more recently, analysing the stylometric ‘fingerprint’ of the document to highlight anomalies in the writing style of the submitting student. ‘Clarify’ does not use this approach, but instead unpacks the document to its component parts, exploring the metadata within and in particular, the edit mark-up or revision save identifiers (RSIDs) that remain within the metadata to help build a picture of the way the document has been created. This in turn allows the assessor to establish whether the document has been created in an authentic manner, with an extended pattern of editing throughout the document, or whether it falls outside this norm by having large sections of unedited text. Other digital artefacts are also examined including font metadata, image information, editing time and more, and these details are reported back to the assessor via a web-based interface. The author has published a number of papers as a result of this work, one on student perceptions of academic integrity, two relating to digital forensics which detail the approaches used in developing ‘Clarify’, and a further two (one published and one in print) on the use of digital forensics and cyber security techniques to assist in the prevention and detection of academic misconduct, all of which have international reach. In particular, ‘Clarify’ as a tool has received international interest and as such is an important contribution to the field of academic integrity.
# Table of Contents

Understanding academic integrity and plagiarism in the digital age: Can digital forensics techniques help prevent and detect academic misconduct? .................................................................i

List of Figures ........................................................................................................................................ viii

List of Tables ......................................................................................................................................... xi

Chapter 1: Understanding academic integrity and plagiarism in the digital age: Can digital forensics techniques help prevent and detect academic misconduct? .....................1

1.1 Introduction ........................................................................................................................................ 1
1.2 A brief history of academic integrity ............................................................................................... 2
1.3 Terminology ....................................................................................................................................... 3
1.4 The impact of technology on academic misconduct ....................................................................... 7
1.5 Development of research focus ....................................................................................................... 8
1.5.1 Adopting a two phased approach ............................................................................................... 9
1.5.2 Overview of Phase 1 - prevention and penalty .........................................................................10
1.5.3 Overview of phase 2 - detection .................................................................................................12
1.6 Research questions ..........................................................................................................................13
1.7 Ethics ...............................................................................................................................................14
1.8 Researcher profile .............................................................................................................................14
1.9 Researcher’s role and contribution to knowledge ...........................................................................15
1.10 Thesis outline ................................................................................................................................16
1.11 Summary of Chapter 1 ....................................................................................................................17

Chapter 2: Methodology .......................................................................................................................18

2.1 Introduction .......................................................................................................................................18
2.2 Philosophical considerations ..........................................................................................................18
2.3 Rationale for research design .........................................................................................................19
2.4 Student Voice Taxonomy ...............................................................................................................20
2.5 Primary research overview ..........................................................................................................21
2.5.1 Relevant research methodologies ..........................................................................................21
2.5.2 Summary of research methodology for phase 1 .......................................................................24
2.5.3 Data collection – survey design and dissemination .................................................................25
2.5.4 Informed consent .......................................................................................................................27
2.6 Methodology for phase 2 – development of ‘Clarify’, a software tool to detect academic misconduct ........................................................................................................27
2.7 Ethical considerations .....................................................................................................................30
2.8 Summary of Chapter 2 ..................................................................................................................34

Chapter 3: Literature Review ..............................................................................................................35

3.1 Introduction ....................................................................................................................................35
Chapter 5: Results of the Software Development of ‘Clarify’

5.1 OOXXML and the structure of a Word document

5.2 The essay writing process

5.3 Stage 1: Using digital forensic techniques to detect contract cheating

Chapter 4: Results of the Primary Research into Student Perceptions

4.1 Results from the two surveys on student perceptions

4.2 Summary of Chapter 4

Chapter 3: Understanding academic integrity and plagiarism in the digital age

3.2 What are Academic Integrity and Academic Misconduct and why do they matter?

3.3 Legal issues

3.4 Integrity as a writing process

3.5 Understanding student attitudes – The Student Voice

3.6 Detecting academic misconduct

3.7 Cyber security and its relevance to academic integrity

3.8 A forensic approach to academic misconduct detection

Chapter 2: A taxonomy of existing research

2.1 Intrinsic and extrinsic methods of plagiarism detection

2.2 Existing tools and techniques used in plagiarism and contract cheating detection

2.3 Using stylometrics to detect contract cheating

2.4 XML analysis

2.5 The role of detection

Chapter 1: What is ‘academic integrity’ and why do we need it?

1.1 What is ‘academic integrity’ and why do we need it?

1.2 The Daubert Standard and linguistics

1.3 Examples of digital forensics use in other fields

1.4 Application of OOXXML in Copyright disputes

1.5 Application of OOXXML in other contexts

1.6 The role of detection

1.7 Students’ understanding of where to get help

1.8 Students’ awareness of academic misconduct penalties

1.9 Students’ understanding of where to get help
5.3.3 Contracted assignment

5.3.4 Findings from the contracted assignment and next steps

5.4 Stage 2: Testing the method on copy-and-paste plagiarism

5.5 Stage 3: Understanding the forensic artefacts

5.5.1 Investigations on a simple Word document

5.5.2 Investigations to explore the editing sequence

5.5.3 File actions and further editing investigations

5.5.4 Summary of findings from Investigations 8 and 9

5.5.5 Entering text in a single session

5.5.6 Preliminary summary of findings from these investigations

5.6 Further exploration of the unzipped folder

5.6.1 document.xml

5.6.2 core.xml

5.6.3 settings.xml

5.6.4 app.xml

5.7 Stage 4: Testing ‘Clarify’ on student submissions

5.7.1 Revisions

5.7.2 TotalTime

5.7.3 rsidR

5.7.4 rFont

5.7.5 W14 textID

5.7.6 w:shd

5.7.7 w:sz

5.7.8 rsidRoot

5.8 Tool development

5.8.1 Analysis of the document.xml file

5.8.2 Findings from initial testing

5.8.3 Values from app.xml

5.8.4 Visual representations

5.8.5 Limitations

5.9 Summary of Chapter 5

Chapter 6: Discussion and Final Working Prototype of ‘Clarify’
Understanding academic integrity and plagiarism in the digital age

Appendix F: Researcher profile

References

Chapter 7: Conclusions and Future Work

Appendix A: Plagiarism tools (intrinsic and extrinsic)

Appendix B: Greyscale versions of images

Appendix C: Investigations

Appendix D: Table of forensic artefacts found in student submissions

Appendix E: First case of contract cheating studied

Appendix F: Researcher profile
List of Figures

Figure 1: Bloom's Taxonomy from Armstrong (2010) ................................................................. 41
Figure 2: Suitability of detection methods for identifying certain forms of plagiarism, Fig 2 in Foltýnek et al., 2019 ................................................................. 53
Figure 3: Plagiarism Detection Tools – Extrinsic and Intrinsic methods ................................ 55
Figure 4: Extended version of Suitability of detection methods for identifying certain forms of plagiarism, Fig 2 in Foltýnek et al., 2019, with addition of internal XML analysis ........ 57
Figure 5: Properties screen of a Word document ................................................................. 63
Figure 6: Digital Forensics OOXML analysis tool, Fu et al. (2011) ........................................ 67
Figure 7: Respondent numbers for the primary surveys ....................................................... 73
Figure 8: Students' perceptions of plagiarism ................................................................. 74
Figure 9: Students' perceptions of 'academic integrity' ....................................................... 75
Figure 10: Students' perceptions of 'contract cheating' ....................................................... 76
Figure 11: Students' awareness of academic misconduct penalties ..................................... 78
Figure 12: Students' understanding of where to get help ..................................................... 79
Figure 13: A film is a collection of other elements. The Last Jedi (The Walt Disney Company: image subject to copyright) ................................................. 83
Figure 14: Properties screen of a Word document ................................................................. 85
Figure 15: Folder structure of unzipped docx file ................................................................. 85
Figure 16: Decompressed docx file ................................................................................. 86
Figure 17: Output of simple docx file with sentence added .................................................. 89
Figure 18: Basic XML markup showing rsidR editing tags .................................................. 89
Figure 19: Output of simple docx file with a word added .................................................... 90
Figure 20: Basic XML markup showing rsidR editing tags following addition of another word ................................................................. 90
Figure 21: XML code snippet from 'document.xml', Johnson & Davies 2020a ................. 92
Figure 22: XML code snippet from 'document.xml' following an edit, Johnson & Davies 2020a ................................................................. 93
Figure 23: First iteration of the tool to display RSID values (Johnson & Davies, 2020a) ....... 94
Figure 24: Author Work Example Extract showing rsidR values (Johnson & Davies, 2020a) ................................................................. 94
Figure 25: Example of Author Work Example Extract following sharing with a second party (Johnson & Davies, 2020a) ................................................................. 95
Figure 26: Visualisation of the original version of the Author Work Example Extract, written by the first party (Johnson & Davies, 2020a) ................................................................. 96
Figure 27: Visualisation of the copied and edited work of the Author Work Example Extract, as viewed from the second party (Johnson & Davies, 2020a) ................................................................. 96
Figure 28: Visualisation of the Contracted Author's work (colour) (Johnson & Davies, 2020a) .................................................................98
Figure 29: Visualisation of the student submission (colour) (Johnson & Davies, 2020a) ..... 98
Figure 30: Frequency of w:rFonts element in sample files (Johnson & Davies, 2020b)...... 104
Figure 31: Frequency of redundant rPr elements (Johnson & Davies, 2020b) ............... 105
Figure 32: Frequency of w:sz element in sample files (Johnson & Davies, 2020b) .......... 106
Figure 33: Number of XML words in relation to document word count .................. 107
Figure 34: Visual of how rsidR values are built in Investigation 1 ......................... 109
Figure 35: Visual of how rsidR values are built in investigation 2 ......................... 112
Figure 36: Colour-coded visualisation of text showing a normal editing pattern .......... 129
Figure 37: Colour-coded visualisation of text demonstrating an unusual editing pattern ... 130
Figure 38: Colour-coded visualisation of text from sample L4002 ......................... 131
Figure 39: Number of revisions in relation to word count .................................. 131
Figure 40: TotalTime (editing time) in relation to word count ................................ 133
Figure 41: Total number of rsidR revision identifiers in relation to word count ........... 134
Figure 42: Ratio of words to unique rsidR values .............................................. 135
Figure 43: Visual representation of L5002 suggestive of patchwriting ..................... 138
Figure 44: Visual representation of L4005, in class test ..................................... 138
Figure 45: Visual representation of L4004, where a sequence of editing can be seen ...... 138
Figure 46: Visual representation of extract of L7004, showing large blocks of unedited text .......................................................... 139
Figure 47: Clarify folder structure .................................................................... 143
Figure 48: Clarify index page ........................................................................... 144
Figure 49: Clarify output – report created on volunteer document ......................... 145
Figure 50: Clarify output – text output of 10-volunteer_authentic1a.docx colour coded .... 146
Figure 51: Clarify output – colour block output of 10-volunteer_authentic1a.docx ...... 146
Figure 52: Clarify output – extract of rsidR values from 10-volunteer_authentic1a.docx ....147
Figure 53: Clarify output – list of unique rsidR values with number of occurrences ....... 148
Figure 54: Clarify output – preliminary images information .............................. 148
Figure 55: Clarify output – number of words per text run represented as bar code ...... 149
Figure 56: Clarify output – core.xml findings ...................................................... 150
Figure 57: Clarify output – app.xml findings ....................................................... 151
Figure 58: Clarify output – settings.xml .............................................................. 152
Figure 59: Clarify output – image cropping report .............................................. 153
Figure 60: Clarify output - colour visualisation of spelling error Sample 1a ......... 154
Figure 61: Clarify output – spelling error Sample 1a ........................................... 154
Figure 62: Clarify output – spelling error in colour blocks Sample 1a ..........................155
Figure 63: Clarify output – spelling error with amendment Sample 1b ..........................155
Figure 64: Clarify output – colour block visualisation of spelling amended sample 1b ........155
Figure 65: Clarify output – visualisation of amending the spelling manually sample 1c ....155
Figure 66: Clarify output – spelling corrected as typed sample 1d ...............................156
Figure 67: Student submission of fabric ideas .....................................................................157
Figure 68: Internet original of student submitted image showing it has been copied from Pinterest ..........................................................157
Figure 69: Student submission, with the implicit suggestion that the work is the student's own .................................................................158
Figure 70: Internet original of the artwork .........................................................................158
Figure 71: Student submission ..........................................................................................159
Figure 72: Original image from the Internet ......................................................................159
Figure 73: Student submission ..........................................................................................160
Figure 74: Original Internet version .................................................................160
Figure 75: Student submission citing the source as Pew Research Centre .........................161
Figure 76: Uncropped version of graphic showing a filename [redacted] dated 2019 .........161
Figure 77: Random_white_space as viewed by the reader ..............................................162
Figure 78: Clarify output - white text report ................................................................162
Figure 79: Clarify output - Text output where hidden letters are revealed .......................163
Figure 80: Clarify output - Copy-and-paste sample, original document .........................163
Figure 81: Clarify output - Copy-and-paste sample, copied document .........................164
Figure 82: Clarify output - submission found guilty of contract cheating .......................166
Figure 83: Clarify visualisation showing punctuation edits ..............................................168
List of Tables

Table 1: Summary of a selection of key academic integrity surveys ........................................45
Table 2: A Defence in Depth approach to academic misconduct, from Johnson et al. (2022)60
Table 3: Structural elements in an OOXML document ..............................................................87
Table 4: Attribute description, Johnson & Davies 2020a ...........................................................92
Table 5: Attribute description including additional value (Johnson & Davies, 2020a) ........93
Table 6: Author Work Example Extract ratio of edits (Johnson & Davies, 2020a) ...............95
Table 7: Author Work Example Extract ratio of edits after sharing and editing (Johnson &
Davies, 2020a) ..........................................................................................................................96
Table 8: Comparison of Contracted Author and Student Submission (Johnson & Davies,
2020a) ........................................................................................................................................97
Table 9: Full document comparison of ratios (Johnson & Davies, 2020a) ............................98
Table 10: Document and devices (Johnson & Davies, 2020b) ..............................................100
Table 11: Documents created and device information (Johnson & Davies, 2020b) ........102
Table 12: Observations from investigation 1 ..................................................................................110
Table 13: Observations from investigation 2 ..................................................................................113
Table 14: Summary of findings from the investigations .............................................................120
Table 15: Word count results from app.xml ............................................................................123
Table 16: Word count results from app.xml on longer samples ...........................................123
Table 17: Word count results when checked before saving .....................................................124
Table 18: Student submissions for testing ...................................................................................125
Table 19: Artefacts recorded for sample documents .................................................................126
Table 20: Percentage of revisions in relation to word count ..................................................132
Table 21: Ratio of words to rsidR values ...................................................................................136
Table 22: Essay Mill and Student Submission comparison ........................................................167
Table 23: Comparison word ratios between authentic work and academic misconduct samples
................................................................................................................................................171
Table 24: Summary of features provided by 'Clarify' ...............................................................176
Chapter 1: Understanding academic integrity and plagiarism in the digital age: Can digital forensics techniques help prevent and detect academic misconduct?

1.1 Introduction

The digital age brings with it significant opportunities for academia, in connecting people and ideas and supporting collaboration across continents. Alongside these opportunities come a range of challenges, especially in relation to academic integrity, as technology can facilitate new ways for students to engage in academic misconduct. This thesis takes a combined approach to exploring the integrity and motivations of students and their perceptions of academic integrity, and the response of Higher Education Institutions to the challenges of academic misconduct. A new method of detecting academic misconduct will be proposed, using similar technologies to those used in the field of digital forensics, an area that has not yet been applied to academic misconduct.

Academic integrity is of vital importance in education because without integrity the value of any qualification awarded is brought into question. Yet plagiarism, contract cheating and impersonation continue to be significant challenges for education providers. Assuring that the work used in assessment of these qualifications was created by the student submitting it is essential, yet increasingly difficult to do. Existing studies into academic integrity come from a variety of foci, including explorations into the drivers that lead students to engage in academic misconduct (Bretag et al., 2014; McCabe et al., 2012), development of cultures and strategies that help prevent and manage misconduct (Christensen Hughes & McCabe, 2006; Lancaster, 2022a), different cultural approaches to academic integrity (Foltýnek & Glendinning, 2015; Michalska, 2014), and development of policies that provide frameworks for the appropriate handling and management of misconduct through penalties and study skills (Ellis, 2022). Finally, there are studies focusing on the detection of academic misconduct, both by developing support mechanisms and guidance for assessors (Crockett, 2022; Rogerson, 2017), as well on plagiarism and contract cheating detection tools (Chowdhury & Bhattacharyya, 2018; Foltýnek et al., 2020). There are similarities, too, between the way academic integrity can be approached and the role of cyber security in reducing opportunity, detecting and evidencing academic misconduct, which has received early discussions from Dawson (2020), Johnson et al., (2023, in print), Lancaster (2022b),
Murdoch and House (2019), all of whom propose that technological approaches to cyber security can be relevant and beneficial in dealing with the challenges created by academic misconduct.

1.2 A brief history of academic integrity

Academic integrity is important in academia because Higher Education Institutions (HEIs) award qualifications based on a student’s overall command of their subject, as measured against standard benchmarks. Such qualifications assure the 'competency' of the student in a specified area, and an important aspect of that is confidence that the institution has been evaluated externally as being in a position to qualify its graduates for employment in a particular field (Woodhouse, 1999). If the work assessed is not that of the student, then this quality assurance is not guaranteed, bringing the reputation of the awarding HEI into question. Reputational damage is not the only concern though. The UK’s Quality Assurance Agency (QAA) suggests that there are significant risks if a student gains an award that is not based upon their own work (as in the case of both plagiarism and contract cheating), noting the real risk that it could endanger public health and safety (QAA, 2022). Renowned academic in the field of academic integrity, Tracey Bretag, stated that plagiarism and contract cheating are serious academic issues that ‘undermine the integrity of education’ (Bretag, 2013).

Discussions around academic integrity often begin with plagiarism – the notion of copying the work of others and passing it off as your own. This is a not a new concept. One of its earliest appearances is from the Roman poet Martial, who used the word ‘plagiarus’ to describe performers who passed off his work as their own (Bailey, 2011), although the term ‘plagiarism’ appears in the dictionary for the first time in 1755. Nor is plagiarism restricted to just academia. In an article in The Guardian, Moss (2005) lists a whole host of works that plagiarise the work of others, including Samuel Taylor Coleridge, Martin Luther King, and George Harrison, and notes that T S Eliot’s ‘The Wasteland’ pays homage to Shakespeare, Chaucer, Webster and others, calling this a kind of ‘verbal kleptomania’. Plagiarism is not always covert either – it is noted from several sources that when Brahms was accused of having used a theme of Beethoven’s in his own compositions, he simply agreed that “any fool could see that”, though references to this differ on which works of Beethoven’s were copied. A more recent examples of plagiarism in the arts is that of Ed Sheeran, who was accused of plagiarising Chokri and O’Donaghue’s 2015 song ‘Oh Why’ in his 2022 hit single
‘Shape of You’, but the case was not upheld, and Sheeran was awarded damages worth over a million dollars (Zemler, 2022).

Plagiarism can also be seen outside of the arts and literature fields. In her National Convention speech of July 2016, Melania Trump’s words bear some striking similarities to a speech given by Michelle Obama in 2008 (Allen, 2016), and her daughter Ivanka was accused of copying almost every detail of a shoe called “Wild Thing” designed by Italian shoe company Aquazzura Italia, though the case was later dismissed (Stempel, 2017).

1.3 Terminology

Terminology evolves as understanding around the topic develops. In academia, education providers are moving away from the accusatory language associated with academic misconduct (‘cheating’, ‘copying’, ‘plagiarising’) towards a greater emphasis on the importance of ‘integrity’; the focus and language has evolved, with plagiarism and academic misconduct being the punishable, darker side of academic integrity and with greater emphasis on students maintaining integrity in their studies. It is therefore useful to define what is meant by ‘integrity’, to consider what this means in the context of academia and education, and to discuss if and why this is significant. According to the Cambridge Dictionary, integrity is “the quality of being honest and having strong moral principles that you refuse to change”.

The website Plagiarism.org (plagiarism.org, 2017), defines plagiarism is an act of fraud, which involves ‘both stealing someone else’s work and lying about it afterwards’ – in other words, submitting work as your own when it has actually been taken from another source. The European Network for Academic Integrity (ENAI, 2022) defines integrity as “Compliance with ethical and professional principles, standards, practices and consistent system of values, that serves as guidance for making decisions and taking actions.” How this applies in an academic sense is of considerable importance to this research – what is acceptable for one person may not be acceptable to another; what is acceptable in one culture may be considered wholly inappropriate in another. To further complicate matters, individual institutions will have their own take on the meaning of academic integrity. For example, the author’s institution describes academic integrity as ‘being honest and sticking to morals that most people would agree with’, which has a much more general and informal tone, though it goes on to mention the importance of maintaining high standards of academic conduct and professional relationships (University of South Wales, 2022).
Whilst academic integrity might, in the first instance, seem like a relatively easy topic to address, it is, in fact, very complex, interdisciplinary in nature, and in need of a diverse approach (Bretag, 2016). Research exists with a variety of foci, including research into the most appropriate definition of academic integrity, studies on the different types of academic misconduct, and many discussions on why students engage in academic misconduct. All these areas are important and relevant. Bretag (2016) notes that academic integrity is a multifarious topic, with interpretations of the meaning of the term differing across the globe. Fishman (2016, p. 7) discusses how academic integrity discourse is often framed as morally and ethically wrong, that it may be framed in a criminal / illegal context, and even that it brings language consistent with disease, such as ‘contagion’, ‘fester’ and ‘rampant’ (McCabe et al., 2012). Park (2003) observes that there are seven themes relating to plagiarism: meaning and context, forms of plagiarism, reasons for plagiarising, how extensive a problem it is, plagiarism from a digital perspective, why academic integrity should be promoted, and finally, student views on plagiarism. Foltýnek et al. (2020) provide an excellent overview of the different types of plagiarism, categorising them into the following groups: character-preserving plagiarism (such as copy-and-paste); syntax-preserving plagiarism (which involves using the same information but with synonym substitution or technical disguise); semantics-preserving plagiarism (which includes translations and paraphrasing); idea-preserving plagiarism (which uses concepts and ideas) and finally, ghostwriting (or contract cheating, whether for financial benefit or otherwise). Clarke and Lancaster (2006) first use the phrase ‘contract cheating’ to describe a process whereby a student will submit original work as their own, despite the work being written for them but not by them. The term is fleshed out by Draper et al., (2017), who add comments around the options for students to specify grades, referencing styles and drafts from these authoring services, as well as explaining how these essay writing services target students.

Students may have a different view of what constitutes academic integrity to that of their assessors. A single person may interpret ‘integrity’ differently depending on the circumstances. Indeed, it could be considered that there are ‘degrees of integrity’ – is one person’s view of integrity the same in every circumstance they encounter, or could it be acceptable to have differing standards depending on the situation? Even within academic institutions, the term ‘integrity’ can be open to interpretation, and assessors will make judgements about the integrity of work submitted by students depending on their own views of academic misconduct and integrity and what is and is not acceptable.
What happens when academic integrity is not upheld will vary from person to person and from institution to institution. As the ‘naughty sibling’ of academic integrity, academic misconduct can fall into several different categories. The author’s institution lists the different types of academic misconduct as below, though it should be noted that other types of misconduct exist including impersonation, bribery, deception, fabrication, patchwriting and more:

- Plagiarism
- Cheating
- Contract Cheating
- Falsification
- Recycling and
- Collusion

Fishman (2009) proposed a new definition of plagiarism because she believed that at the time existing definitions around plagiarism focused too heavily on laws relating to theft and copyright. This definition is very clear, focusing on the act itself and the context, and forms an excellent underpinning definition for this thesis. Fishman claims that plagiarism does not equal theft, and nor is it the same as copyright violation, but instead suggested that plagiarism needs to be identified by five constituent elements. She homes in on the notion that plagiarism is an offence not because of the taking per se, or the loss of tangible goods to the original owner, but because the student is asking for recognition and educational gain for work that has not been duly earned, and this is the underpinning reason why prevention and detection are so important in the education system. Fishman’s proposal (Fishman, 2009) is that plagiarism occurs when someone:

1. Uses words, ideas, or work products
2. Attributable to another identifiable person or source
3. Without attributing the word to the source from which it was obtained
4. In a situation in which there is a legitimate expectation of original authorship
5. In order to obtain some benefit, credit, or gain which need not be monetary
In terms of contract cheating, described by Clarke and Lancaster (2006) as ‘the process through which students can have original work provided for them, which they can then submit as if this were their own work’ each of the five statements given by Fishman still hold true and as such form an excellent benchmark for this research study.

The areas of particular interest for this thesis are plagiarism, contract cheating and to a lesser extent collusion, which are defined by the author’s institution (University of South Wales, 2022) as follows:

**Plagiarism** is when students take someone else’s work or ideas and pass them off as their own. Plagiarism may be written or non-written.

**Contract cheating** is when students seek other persons to produce work or buy an essay or assignment, either ‘off the shelf’ or specifically written for them, and submit it as their own work.

**Collusion** occurs when, unless with official approval (e.g. in the case of some forms of group projects), two or more students collaborate in the production of work, and they submit work which is unreasonably similar and/or is represented by the students to be the product of their individual efforts. One student sharing their own work with another student, resulting in similar pieces of work being submitted, is considered to be collusion by both parties regardless of intent.

Of course, plagiarism, contract cheating and collusion are not confined to one subject or indeed, one field, they are transdisciplinary in nature. The term ‘transdisciplinary’ is generally attributed to Piaget (Bernstein, 2015). Bernstein describes transdisciplinarity as being ‘characterized by its focus on “wicked problems” that need creative solutions’. Whilst some academic subject areas may appear to have more cases of cheating, there is limited evidence to support this, and similarly there is no one ‘type’ of student that will engage in academic misconduct, nor one type of institution or level of study. As such, investigating academic integrity and academic misconduct can be very challenging, as can finding appropriate deterrents and methods of detection. The transdisciplinary nature of academic integrity will feature later in discussions, as approaches to dealing with it can be borrowed from other areas such as cyber security.
Academic discussions on these topics are important, but gathering students’ views on academic integrity and academic misconduct is also key to understanding why misconduct occurs and what can be done about it, as this forms part of a dual approach to dealing with this complex issue. As a vast and complex topic, spanning multiple disciplines and affecting students and academics alike, this thesis will focus on several of these areas, starting with student perceptions of academic integrity, and then moving into exploration around the detection methods currently used and potential new ways of detecting academic misconduct.

1.4 The impact of technology on academic misconduct

Technology has a significant role in this field of research. It can facilitate academic misconduct by making it easier to copy-and-paste materials from sources that are readily available on the Internet, through library databases, websites and journal repositories. The ability to find and re-use other people’s work became much easier during the 1990s because of the simplicity of copy-and-paste (introduced in 1983) and the availability of basic search facilities (Rettinger & Bertram Gallant, Ed., 2022). Technology also makes it much easier (and less ‘personal’) to contract work out to another author through the use of file sharing websites and essay writing services. Contracted authors and people impersonating students for online assessments can be hundreds if not thousands of miles away, on different continents, yet still providing work for students wherever they may be, day and night.

Very recently, there has been a big increase in the availability and visibility of artificial intelligence for writing essay answers. Users are able to ask an artificial intelligence ‘bot’ (e.g. ChatGPT, OpenAI) a question, giving it specific instructions such as ‘write a 1000 report on xxx, including an introduction, two or three key points and a conclusion. Include relevant theories and references’. When the outputs are not quite as intended, it is possible to fine tune the results to suit. This brings a huge challenge to the academic world in ensuring that students who submit work using tools such as this fully understand the topic they are being assessed on, and are appropriately assessed with this in mind.

Yet whilst technology can make academic misconduct easier, it can also help provide solutions to these complex issues, at least in part. Text-matching software has been used for some time to help identify passages of text that have previously been used by looking for phrases and passages that match text in on and off-line repositories. Tools such as Turnitin (https://www.turnitin.com), PlagScan (https://www.plagscan.com) and Urkund
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

(https://www.urkund.com) are excellent at carrying out such text-matching. It is also possible for an assessor themselves to copy passages of text and paste them into a Google search (or similar) with quotes surrounding the passage to see if there are any immediate online matches. More recently ‘authorship tools’ have been developed. These use stylometry and linguistics, as well as exploring some of the document metadata, to determine matches between authors (whether the submitting author, or a third party), and to search for anomalies in submissions by a single student (suggesting that one or more assessments may have been outsourced). Whilst these two combined approaches provide very useful information, they focus primarily on the textual content of the document submitted, and do not explore the document writing process itself to establish authenticity. This is a particularly important area to understand for detection of work generated by essay mills or artificial intelligence.

Other techniques take a different approach to using technology to help prevent and detect academic misconduct. It is possible, for example, to deploy cyber security approaches in an attempt to prevent academic misconduct, by blocking network access to essay mill sites and sending a pop-up screen to a student to suggest that they might benefit from speaking to a study skills team. Cyber security system monitoring can also be used to track student log ins (Murdoch & House, 2019), flagging any unusual behaviours that could indicate impersonation. Finally, digital forensics techniques can be repurposed to detect forensic artefacts in a student submission that help build a picture of how a document has been created (Johnson et al., 2022). All of these novel techniques are in their infancy, providing interesting avenues to explore and encouraging a holistic approach to academic integrity and misconduct in a way that mirrors technological advances, digital transformation, and cyber security.

1.5 Development of research focus

According to Foltýnek et al. (2019), there are three main considerations when it comes to academic misconduct, all of which are important components of the academic integrity landscape. These are prevention, detection and punishment. Prevention considers policies, workshops and interventions that may help support students during their education, thus giving them sufficient tools and networks that deter them from considering any form of academic misconduct at key pressure points during their studies. Prevention is unlikely to be fully achieved, and Lancaster & Clarke (2016) suggests that for contract cheating in particular, it can be helpful to think of both prevention and detection as linked concepts. Detection includes support and education for staff in how to spot academic misconduct and
what to do when they find it (QAA, 2022a; Rogerson, 2017), as well as implementation of appropriate technologies to help with this detection (Lancaster & Clarke, 2016, Lancaster, 2022b; Weber-Wulff, 2016). There is a societal impact to allowing academic misconduct to go undetected, potentially putting people with unearned qualifications into roles that put people’s lives at risk. Imagine, for example, someone awarded a medical degree who goes on to operate on patients without fully understanding the consequences of their actions, or an engineer who is unable to carry out mechanical calculations effectively, even though he/she has an engineering degree, and their employers believe they have the necessary skills to carry out such work. And there are similarly serious implications for people who qualify in law, business, nursing, education and so on. Furthermore, awarding institutions want to ensure that students who do complete the work on their own are duly rewarded for their efforts, and that honest students are rewarded fairly. Punishment focuses on the equitable and appropriate penalties given to students who engage in academic misconduct. Whilst the substantive part of this thesis focuses on the detection side of academic misconduct, it is important from an ethical and moral perspective to consider what may lead a student away from academic integrity. The thesis therefore takes a two phased approach which is described below.

1.5.1 Adopting a two phased approach

In the initial phase of the research, a review of literature relating to student perceptions of academic integrity and academic misconduct was carried out. This helped to build a taxonomy of existing research in this area, and was followed by primary research using surveys at the researcher’s institution to establish whether results echoed existing research. The surveys were carried out at two points in the year to enable reflection on any changes in student perceptions during this time. Whilst the initial intention was to continue this research in more depth using focus groups, the researcher’s growing interest and expertise in cyber security, along with an incident of contract cheating that presented itself at a timely point in the study, meant that another area of interest was developed. This focused on detection of academic misconduct through exploration of the metadata that exists in Word documents, and early findings yielded very interesting results that were worthy of further exploration. This evolved into the second phase of the research, which was to develop a tool to explore the Office Open Extensible Markup Language (OOXML) format, analysing the construction of a Word document, and returning the results in a visual format to the assessor.
The final output of the thesis is a working prototype of an academic misconduct detection tool with the working title ‘Clarify’. This tool looks at the metadata of a given document, exploring the forensic artefacts that accumulate as the document is written and constructed. It highlights artefacts that may be indicative of academic misconduct providing a quick and automated way to analyse student work. The tool is presented along with examples of the information it can provide, and possible future developments.

1.5.2 Overview of Phase 1 - prevention and penalty

Developing an understanding of why students may cheat can significantly help raise awareness amongst educators, helping in the development of authentic assessment, support mechanisms and policy. Existing research around why students cheat is explored in more detail later, but some clear and consistent themes emerge. In a systematic review article by Newton (2018), a summary of the main themes as to why students cheat is provided, which include past cheating behaviour, the use of honour codes, poor study conditions, grades (poorly performing students being the most likely to cheat), lack of motivation, ‘normalisation’ of cheating, studying in a second language, age and many more. Other research suggests that students may engage in academic misconduct because they find themselves on a programme that is too difficult and yet face pressure to achieve (Blum, 2016). Nilson (2010) suggests that the Internet is to blame for increasing ‘cheating’.

Anderman et al. (2022) provide a thorough psychological walkthrough of academic motivation and the relationship between ‘cheating’ and motivation. Curtis (2022) draws together many of these reasons too, adding the increasing use of online learning into the mix alongside dissatisfaction with the learning and teaching environment. The QAA’s Contracting to Cheat report echoes many of these points (QAA, 2022). Chen and Macfarlane (2016) comment on the sense of resignation that cheating is commonplace and cannot be eliminated, and that there is a common practice of copying work as part of a culture of reciprocity.

Unethical behaviours can occur much earlier in the education process too as shown by Khan & Mulani (2020), who reported that children were getting help on school projects and assessments from their parents. These unethical behaviours can continue well into working life (McCabe et al., 2012), impacting on professional integrity, raising questions around when children should start to learn about academic integrity. It also highlights concerns around how well-equipped young people are to carry out research and referencing correctly by the
time they reach further and higher education, whether they understand the implications of misconduct on their future professional lives and where in the education system these issues should be addressed. Questions also arise around the purpose of education itself. Children learn because they have to, school is compulsory, and they have little choice about what they learn. But after compulsory education, students have made a choice to study and have a range of motivations for getting higher qualifications, which may not just be about the pursuit of education, but may be because an employer has required a certain qualification, or because a promotion depends on it. The different reasons for study undoubtedly impact on a student’s motivation to do well and to learn with integrity, and Bertram Gallant (2016) talks about the importance of positioning integrity as central to the system of education, rather than focusing on the prevention and punishment of misconduct. Where students do not value their education, or cannot see purpose in it, they are far less motivated to study and may find themselves taking short cuts or leaving work until the last moment (see 3.5). Finally, the question as to whether students really understand why they should reference properly is important, for if they cannot see the value in research and referencing then they are less likely to take care in doing this with integrity.

As mentioned previously, academic integrity is a very complex area, where many elements combine, including students’ understanding of academic misconduct, appreciating the pressures students face when studying and the consequences to them personally of failing any assignments (and how this impacts their decisions), and considering why integrity should matter to students at all. The primary research conducted aimed to see if these findings were reflective of the situation at the researcher’s institution to establish whether the landscape is similar to existing research. Understanding students’ perceptions can greatly help shape the ways that HEIs work with students to reduce the likelihood of them engaging in academic misconduct and it was useful to consider student views on studying for their qualifications, as well as their understanding of the terms “academic integrity” and “academic misconduct”.

This led to initial questions around the following areas:

- What training and support is given to students to ensure they understand what academic integrity is?
- What situations and circumstances lead students to ‘cheat’?
Whilst there are measures that can be put in place to help reduce academic misconduct, it would be foolhardy to suggest that it can be eradicated altogether. Many papers describe techniques that may help reduce the opportunities to plagiarise and outsource work and it has been suggested that authentic assessment can help prevent contract cheating in particular, perhaps through using ‘real-world’ tasks (Ellis, 2020). Rogerson (2017) suggests that appropriate training can help assessors detect contract cheating by teaching them the signs to look out for, such as citation and referencing irregularities, inconsistent language, inconsistent writing styles within a single submission and across a portfolio of student work. She also notes the importance of ensuring that assessment tasks are refreshed regularly and cross-checked on the Internet to ensure the questions do not appear on file sharing and essay answer websites.

Both prevention and penalty are deterrents to students and in an ideal world, academic misconduct simply would not happen because these deterrents would be sufficient. Students would fully grasp the notion of academic integrity, they would value the time and effort that goes into writing original work, and they would complete their education accordingly. This is simply not realistic, and given that it is not possible (or at least highly unlikely) to be able to completely prevent academic misconduct from occurring, HEIs (and indeed, all awarding bodies) need to be confident that the student who is submitting work for credit is the genuine author. This highlights the need for robust, reliable and effective plagiarism detection methods.

1.5.3 Overview of phase 2 - detection

The use of text-matching software as the primary method for detecting academic misconduct has already been mentioned but this has its limitations, which will be explored in greater detail later in the thesis. Text-matching software compares submitted student work with online web resources, using search engines or offline database resources (Foltýnek et al., 2020), and provides a good starting point for detection of cut-and-paste plagiarism. More recently, work has been done on developing authorship tools, which examine the stylistic attributes of a document to look for anomalies in the way the work is written, which can help detect contract cheating (where the work may indeed be original, albeit not by the submitting student). An alternative or additional approach to detection could be to explore the application of cyber security tools and methodologies, and in particular around investigative techniques relating to cyber-crime. Repurposing digital forensics techniques could have the
potential to provide an alternative and novel method of detection, by exploring the inner workings of a document to look for much more detailed forensic artefacts. In addition, the application of cyber security approaches to academic integrity and academic misconduct seem to present themselves as a very appropriate way to strengthen institutional defences and approaches to tackling misconduct, given the transdisciplinary nature of both challenges. These tools are useful, but do not provide a ‘silver bullet’. Instead, they provide ‘flags’ that let the assessor know that things do not seem quite right and further investigation is warranted.

Digital forensic investigation techniques are used primarily in law enforcement to prove ownership of a particular digital device that may have been involved in a crime, or to provide evidence of a suspect’s involvement in a crime. In addition, digital forensics techniques have been used in cases relating to the theft of intellectual property (Jeong & Lee, 2017) to try and establish the true or original ownership of digital documents. In Jeong and Lee’s case study, a hypothetical case was created where a trade secret was improperly acquired. Analysing the computer where the trade secret was initially created and the USB containing the alleged copy using digital forensics showed that the documents were indeed linked, by reviewing revision history and temporary files. However, in this example, the analysis was dependent on having access to the devices on which the trade secret was stored. Nonetheless, an interesting application of similar techniques is in the detection of academic misconduct, as it may be possible to identify the original source of materials within a student submission from its digital fingerprint, or to establish unusual patterns of editing and forensic artefacts that could be indicative of contracted out work.

1.6 Research questions

This thesis focuses on developing a novel method of detecting academic misconduct by exploring the fields of cyber security and digital forensics and how techniques developed in these disciplines can be adapted for use in detecting plagiarism and contract cheating. However, no exploration of detection techniques would be complete without taking a holistic and ethical approach to understanding what leads students into cheating behaviours, and as such, a two phased approach was taken to firstly gather primary data directly from students, and secondly to explore the use of digital forensics to detect academic misconduct alongside development of a novel detection tool using these approaches.
The two key aims to this project are therefore as follows:

1. To understand student perspectives towards academic integrity
2. To explore whether detection of plagiarism and contract cheating can be improved using digital forensics techniques

The research questions are stated as:

1.a What is the focus of existing research into academic integrity / misconduct?
1.b What types of data already exist?
1.c To what extent is the student voice heard in existing research?
1.d How will hearing the student voice help in reducing plagiarism, and in particular contract cheating?

2.a What tools currently exist to help detect plagiarism and contract cheating?
2.b What digital forensics tools (if any) are used in other areas that could be applied to plagiarism and contract cheating?
2.c What is the writing process, and can this be reverse-engineered to find out more about the authenticity of a student submission?
2.d To what extent is it possible to develop digital forensics techniques to help examine the essay writing process and therefore identify plagiarism and contract cheating?

1.7 Ethics

All activities that took place during this research project were undertaken with due consideration to ethics. Discussions took place with different people at the researcher’s HEI at various times during the research. This was to ensure ethical approval was granted for each stage, as well as to discuss the most appropriate approaches for working with students during both phases. There were four main points where ethical approval was sought and obtained, and these are detailed in section 2.7. The ethical framework is discussed in full detail in the Methodology.

1.8 Researcher profile

The researcher first became interested in academic misconduct whilst teaching information technology at a further education institution. A number of students submitted work which
was clearly plagiarised, and yet did not expect to get caught, nor appreciate that what they had done was plagiarism when confronted. This led the researcher to consider aspects around student perceptions of academic integrity, what they valued about education and what motivated them to engage in plagiarism. When the researcher moved into higher education, the problems were compounded as awareness around contract cheating grew and a broader interest in different types of academic misconduct as well as around detection of these various types developed. Early discussions with the researcher’s supervisor encouraged consideration of student motivations, rather than from a punitive ‘we must catch them’ perspective. Also, during the early part of the study, the researcher began teaching cyber security and digital forensics, and knowledge and expertise in this area developed. Having a background in the arts, the researcher was not a fully trained software developer and had only self-taught programming skills, but could see potential in repurposing some of the digital forensics techniques taught in cyber security classes for reverse-engineering written documents. This accounts for the less formal approach to software development discussed in Chapter 4. This less traditional route into cyber security and software development is considered advantageous however, as it meant the researcher was able to view academic integrity and academic misconduct through a wider lens, approaching it from a holistic viewpoint and not being constrained by traditional approaches to academic misconduct and its detection. A full profile of the author is provided in Appendix F: Researcher profile.

1.9 Researcher’s role and contribution to knowledge

Unlike existing detection methods, the software developed during this research fully unpacks the submitted document in a similar way to reverse-engineering techniques in cyber security, and explores the editing process that underpins the writing, comparing it to what can be expected from an authentically written document.

At the time of completion, the researcher has had four peer reviewed articles / book chapters published with a fifth in print. These include a book chapter on student perceptions; two journal articles and a book chapter on the application of digital forensics techniques to academic misconduct detection; and a third book chapter on using cyber security approaches to academic integrity. The researcher was lead author on all these publications, with supporting authors, who were on the PhD supervision team, providing discussion around the topics, proof reading and grammar checks. Full details of all publications can be found in Appendix F: Researcher profile and Appendix I: Published works.
In terms of the software development, preliminary discussions and testing was carried out with the researcher and a supervisor to explore ideas. Following this, all tool development was the responsibility of the researcher with support from supervisors only given for particularly challenging problems. In many cases, these problems were resolved by the researcher being directed to appropriate resources to assist with learning and development.

There has been international interest in ‘Clarify’ and the methods it uses to highlight potential academic misconduct. This includes requests to access the tool from renowned researchers in plagiarism and contract cheating as well as requests to read this PhD thesis. Details of these requests can be found in Appendix F: Researcher profile.

1.10 Thesis outline

This thesis examines the topic of academic integrity and academic misconduct, with a particular focus on two key areas, firstly, that of student perceptions towards academic integrity and academic misconduct, and secondly, looking at how academic misconduct is detected, with a particular aim of developing a new detection method. The two phased approach reflects the researcher’s desire to take a holistic approach to academic integrity and is steered towards the development of ‘Clarify’ because of the researcher’s background in cyber security, an area which shows promise in terms of additional methods of detection.

The structure is as follows:

1. Introduction, introducing the research topic, key terms, research aims and questions and planned outputs from the research.

2. Methodology, including the approach used to establish a taxonomy of research into the student voice and rationale for methods chosen for the primary research into student perceptions at a single institution, followed by the software development methods used and testing regime for ‘Clarify’, exploring the digital artefacts available in a document and developing a tool to capture this information.

3. Literature Review, exploring existing research into student perceptions and identified gaps, exploration of current academic misconduct detection tools and their limitations, cyber security approaches in Higher Education Institutions and relevance to academic integrity and detection of academic misconduct. During this chapter the OOXML format is explained in more detail, and the approach used in development of the software to
detect academic misconduct is discussed alongside various worked examples of its potential use.

4. Results from phase 1: student perceptions.

5. Results from phase 2: development and testing of ‘Clarify’.

6. Discussion, including an evaluation of final working prototype of ‘Clarify’, along with sample files to demonstrate its function.

7. Conclusion and recommendations for future work

1.11 Summary of Chapter 1

This chapter has introduced the two complementary aims of this thesis, which are to explore students’ perceptions of academic integrity and to investigate the application of cyber security techniques, and specifically digital forensics, to academic misconduct detection. The terms used in the research were discussed, along with an overview of the key areas of academic integrity, academic misconduct, student perceptions, and cyber security in relation to academic integrity. An introduction to the dominant themes around academic integrity was provided. The research questions were presented and followed by an overview of the thesis structure. The next chapter will discuss the methodology and methods used within this research.
Chapter 2: Methodology

2.1 Introduction

This chapter will discuss the two phased approach to the research, outlining the methodologies considered and rationale for selections made.

2.2 Philosophical considerations

In any research, the methodology used will depend on the nature of the subject being researched and will be influenced by the researcher’s ontological (nature of reality) and epistemological (nature of knowledge) standpoints. Because of the depth and breadth of issues associated with academic integrity and its value in the learning process, it is particularly important to consider these philosophical aspects as they will help to inform the most suitable methods of data collection and investigation, analysis and evaluation (Bryman, 2016). Whilst students and faculty may all consider academic integrity to be important, they do not, perhaps, consider what it really means. Indeed, looking at literature from other countries, and from a transhistorical lens, it is clear that there are many variations on the understanding of the word integrity and its importance, or otherwise, in academia. In Chapter 1, several definitions of integrity were offered, including that of ENAI (2022), who define integrity as “Compliance with ethical and professional principles, standards, practices and consistent system of values, that serves as guidance for making decisions and taking actions”.

Consideration was also given to differing views of integrity between cultures, between people from each culture, and even within one person when applied to different contexts.

Ontology considers the nature of social phenomenon and how the researcher understands this, posing the question of whether social phenomena are something beyond our control, or are a produce of social interaction (Bryman, 2016). An objectivist approach asserts that social phenomena and their meanings are independent of social actors, and that they are beyond influence. In the case of academic integrity, this would suggest that academic integrity is a thing that cannot be changed or influenced by others, or by individual thought or the community that surrounds it. Most people are likely to hold the inherent belief that integrity is ‘good’ and to be sought after, but beyond that it would be foolhardy to consider that a realist ontological approach – i.e. that any one truth about the meaning of integrity and specifically academic integrity – is the most appropriate. The alternative ontological
approach is that of constructionism, which asserts that social phenomena are constantly being revised by social interaction, and in the case of academic integrity, this means that developments in technology, the culture of an organisation, even the self-development of individuals throughout their studies will alter the viewpoint. The challenges of what constitutes academic misconduct are further compounded when considered from an epistemological standpoint. Contextual understanding helps to highlight the issues of cultural differences, whilst subjectivist epistemology leads to the view that each individual will have their own understanding of the topic. All these factors highlight the needs for educational institutions and assessors to be very clear with their students concerning what is and is not acceptable in terms of academic integrity.

It is highly beneficial to understand students’ perceptions of academic integrity, to explore the extent to which students understand what academic misconduct is, and the associated penalties and consequences.

In summary then, the following complimentary areas will be explored:

1. Student Voice: Overview of existing research into students’ perceptions of academic integrity and misconduct, discussed in Chapter 3, followed by primary research at the researcher’s institution, with results discussed in Chapter 4.

2.3 Rationale for research design

In the first instance, a review of existing research on the student voice in relation to academic integrity was conducted to explore students’ perceptions of academic integrity and academic misconduct. This formed a background literature review and findings from this activity are detailed in Chapter 3. Following this, primary research was carried out at the researcher’s institution to see if the results were consistent with the previously published and analysed research in this area. The primary research took a case study approach of a convenience sample, using surveys as the data collection instrument. Upon completion, an opportunity to analyse a student submission that had been contracted out presented itself. Having established that student perceptions at the researcher’s institution were not dissimilar to previous research
findings, curiosity and the researcher’s expertise in digital forensics meant that the research
naturally evolved to examine this student submission forensically leading to a greater focus
on the detection of academic misconduct. Analysis of this submission highlighted the
potential value of XML metadata analysis of student work, and development of the software
tool commenced, firstly by carrying out various technological investigations and then by
using an iterative approach to software development through more specific investigations.
Work created by the researcher was tested to explore specific features of XML. The findings
at this point were of sufficient interest to warrant testing on student work that had been
through an academic misconduct panel. At all stages, tests were documented in detail, and
manual analysis carried out which mirrored the automated tasks to check for accuracy. An
informal product control list was added to the software to monitor progress and to suggest
feature additions. The final stage was to create documentation which could act as a user guide
to the tool’s various functions.

2.4 Student Voice Taxonomy

The initial literature review carried out on student perceptions created a useful taxonomy
from which to start. Findings from this part of the research were published in Johnson and
Reddy (2018). It involved carrying out a search on the researchers’ university database,
which accessed articles and journals from the following scholarly databases: Academic
Search Complete, Wiley, ERIC, Elsevier, Emerald, ProQuest, Springer and others. A number
of search terms were used including ‘Plagiarism’, ‘Academic Integrity’, ‘Academic
Misconduct’ and ‘Cheating’ using Boolean operators, and with the criteria of ‘peer
reviewed’, and a date range of the past twenty years. This search yielded just under 170,000
results and a ‘subject’ search was subsequently used to refine the results. This uses the data
from the MARC 650 field, based on National Library of Congress subject headings, and
yielded a much more focused set of results of just under 4,200 publications at the time of
writing. No exclusion criteria were used to ensure that all aspects of plagiarism would be
included in the result. A manual review of the results was then carried out and items not
relevant to the search were excluded. These included articles about contract cheating markets
and academic misconduct in the commercial world amongst other things. Once this first sift
had been completed, the abstracts of the remaining articles were read to establish whether the
student voice / student perceptions were mentioned. If so, the article was skim read to look
for information coming directly from students, perhaps through quotes or summaries of
themes, this was done via an initial coding analysis and key themes were identified. Several other documents known to the researcher and supervisors were also reviewed as appropriate.

Alongside peer reviewed journal articles, other sources of potential information relating to the student voice were searched for via an Internet search. A number of these resources were found from a Facebook group dedicated to Academic Integrity to which the author belongs. Once a list of resources had been compiled, they were categorised into the following categories: Surveys; Written discussions and guidelines; and Internet sources.

2.5 Primary research overview

Following this initial literature review on student perceptions, a case study of a single higher education institution was conducted, through two survey points carried out with first year students at the start and end of the academic year. It was decided that a survey at the researcher’s institution would provide a good starting point for this primary research. Reviewing changes or developments in perceptions throughout the academic year was not apparent in the surveys gathered from the literature review, and this was an area that could demonstrate how a student’s perspective may change throughout the academic year. The assumption here was that students may feel differently about whether they may cheat or not when they are faced with potentially stressful deadlines. In addition, it would be useful to compare students’ understanding of key terms associated with cheating (plagiarism, contract cheating, academic integrity). Given that there is an expectation on teaching staff to support academic integrity, does this translate to students?

2.5.1 Relevant research methodologies

As the focus of the research was education, pedagogy, and integrity, several social sciences research methodologies were considered. According to Cohen et al. (2018), research methodologies include both quantitative and qualitative research, and may use mixed methods. Indeed, Cohen et al. (2018) clearly suggest that researchers do not adhere ‘slavishly’ to a single methodology, but that a combination of methodologies may be appropriate.

2.5.1.1. SPICE

Cleyle and Booth (2006) note the importance of information professionals becoming “efficient at identifying, formulating and addressing relevant questions from their own
practice”, and suggest a methodology of research questioning for evidence based practice using the acronym SPICE, which includes the following areas of focus: S (Setting) – Where is the research taking place, and in what context; P (Perspective) – who is the research about?; I (Intervention – phenomenon of Interest) – What is the research for?; C (Comparison) – What else? And E (Evaluation)? – how well did it go, what was the result. This was an interesting model for the current project, with the emphasis on ‘perspective’ – in this case, the student perspective. In terms of ‘intervention’, the research aimed to examine how the student perceives the need (or not) for academic integrity, and whether input from faculty staff can help develop this culture.

2.5.1.2. Itdem’d

In her book ‘Action Research in Teaching and Learning: A Practical Guide to Conducting Pedagogical Research in Universities’, Norton (2019) suggests action research is the process of looking at a problem, modifying practice and contributing to theoretical knowledge. Norton comes up with the acronym “Itdem’d”, where the acronym stands for I (Identify the issue), T (think about methods), D (do the research), E (evaluate), M (modify practice), and D (disseminate). Each aspect is described below with its relevance to this particular project:

- Identify the issue – the issue in this case is academic integrity and students’ perceptions of this.
- Think about the methods – surveys with students at the researcher’s institution will form the main method, in order to hear the student voice directly and then compare it with findings from the initial literature review and development of a taxonomy.
- Do the research – at the time of transfer, the first set of surveys (both at the start and end of the year academic year) have been completed, and the first surveys analysed. The second set of surveys is still to be completed, along with surveys for staff and focus groups.
- Evaluate – Results from surveys and focus groups will be analysed fully.
- Modify practice – Following analysis and evaluation, interventions will be suggested and structures put in place to help support students in their studies.
- Dissemination – will be by way of published articles, guidelines for staff, best practice policy for managers (or similar as appropriate).
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

There were some elements of this approach that had merit in relation to the research aims.

2.5.1.3. Case Study

The case study approach, as defined by Cohen et al. (2018) focuses on observing real people in real situations to ‘Investigate and report the real-life, complex, dynamic and unfolding interaction of events, human relationships and other factors in a unique instance’ (p. 376).

Cohen at al. (2018) put together some useful hallmarks of case studies, and state that ‘case studies exist in their own right as a significant and legitimate research method’. One hallmark in particular was of special interest here in relation to hearing the student voice:

*Focuses on individual actors or groups of actors, and seeks to understand their perception of events (Cohen et al., 2018, p.376)*

Most of the listed hallmarks seemed to apply to this phase of the research project, where the researcher would be discussing academic integrity with students, and would be looking at instances of academic misconduct with a view to understanding why they have occurred, focusing on individual student perceptions. Yin (2009, p. 18) defines a case study as an enquiry that investigates ‘a contemporary phenomenon … within its real-life context’, noting that this addresses the ‘how’ and ‘why’ questions surrounding the phenomenon being investigated. Specifically, Yin’s approach incorporates important contextual conditions. The use of exploration questions would be particularly pertinent – the ‘why?’ do students ‘cheat’, or decide not to adhere to principles of academic integrity, whether deliberately or unwittingly.

2.5.1.4. Action Research

‘Action Research is a useful tool for change and improvement at the local level’ (Cohen et al., 2018, p.440). It is an excellent tool for areas where a problem requires a solution, or a change can result in a positive outcome. Cohen et al. suggest that it is a good method to use in a variety of areas and includes areas of relevance to this study such as teaching methods, learning strategies, attitudes and values. Cohen et al. reference the work of Hopkins (1985) who suggests that there is a personal attempt to understand, improve and reform practice. Developing tools and practices that improve students’ understanding of academic integrity may influence the way teachers work, and the way academic integrity is embedded in the delivery of content, such that it builds a culture of integrity amongst students.
2.5.2 Summary of research methodology for phase 1

Various aspects of the described methodologies had merit when considering the research design. The stages of action research suggested by Cohen et al. (2018) seemed useful, but the focus of the research was less about the intervention and more about an exploration of where students currently are, and what may be done to reduce instances of contract cheating and develop a culture of academic integrity. Therefore, action research was considered less useful in this instance as only the early stages of a true action research methodology would be relevant. Similarly, Itdem’d offered a useful structure for the primary research, but given that its roots are in action research the decision was made that Itdem’d would not be the best approach. Elements of the SPICE model initially appeared to be a good option. Taking an approach to questioning that focuses on context, setting and perspective, this was very relevant to the primary research.

However, it was finally decided that the case study approach would be the most appropriate methodology for a study into academic integrity, which looks at how students, academics and managers approach the issue of academic misconduct and plagiarism. Cohen et al. (2018) say that case study is useful when the researcher has little control over events. Certainly, during the information gathering phase of the research, data gathered will be descriptive and there will be little practical input or control from the researcher. The researcher will be observing the characteristics of an individual unit – in this case a university – with the purpose of probing and analysing the ‘multi-stranded phenomena that constitute the life of the unit, possibly with a view to generalising to the wider population to which that unit belongs (Cohen et al., 2081, p.385) – i.e. the purpose was to seek and understand the views of students regarding academic integrity, academic misconduct and plagiarism. The intention will be to establish a supportive and effective structure in helping students to improve their understanding of academic integrity and therefore minimise the occurrence of cheating behaviours. Gathering views from students on what they considered plagiarism and academic misconduct to be and whether they were aware of the consequences of academic misconduct at their institution via a case study approach seemed to be the most suitable methodology for this part of the research.
2.5.3 Data collection – survey design and dissemination

A key focus of the primary research was to gather students’ views, and for this reason surveys were settled on as the most efficient way to do this, as once designed, they could be issued to a number of students across a range of subject areas and would allow for analysis of findings under themed questions. Yin (2009) comments on the importance of using previous development and discussions of theoretical concepts to guide data collection, and previous surveys by McCabe et al. (1993 - 1999) and the GEMS survey (European Network for Academic Integrity, 2017) were used to develop the survey questions, which included a mixture of Likert scale type questions, check box questions and open questions. Since the surveys in this research project build on the work of others, establishing where there are similarities (if any) echoes much of Yin’s thinking regarding a case study approach. The research also sought to explore the reasons why students may engage in academic misconduct, and as such reflected the contextual aspects described by Yin.

Cohen et al. (2018), also discuss Internet surveys as a research methodology, noting that they are becoming predominant as a way of conducting surveys, due to many reasons including reduced cost, improved speed in distributing, gathering and processing data, volume and many more. One of the key disadvantages of this methodology is that they lack the personal touch from an interviewer, so participants cannot be reassured of the reasons for the survey and the higher level of abandonment, where respondents simply stop completing the survey, or fail to return them.

The planning process was similar to that described by Cohen et al. (2018) with the following approach:

- Outline objectives were defined along with type of information to be collected.
- Existing survey instruments were analysed and formed the basis for the survey design and questions.
- A pilot survey was tested on a few people to check that none of the questions were ambiguous or unclear and that the results returned indicated that the questions had been interpreted correctly.
- Approaches were made to relevant staff at the institution in order to determine which students would be involved.
• Ethical approval was sought once the survey questions and approach had been finalised.

• The surveys were carried out at the two touchpoints indicated below.

• Results were analysed.

The questions were designed to elicit the views of students on their understanding of and attitudes towards academic integrity. The purpose of the surveys was to gather both quantitative data (e.g. proportion of students who admit to having cheated) and qualitative data (what is your understanding of the term ‘contract cheating’). All four faculties in the author’s university were asked to take part, in most cases initially through contact with the faculty Head of Learning, Teaching and Student Experience (HOLTSE). The HOLSTE identified staff who may be willing to take part in the survey. Three of the four faculties were happy to be involved in the study. The fourth was not included due to issues with ethical approval (see sections 2.7 and 6.15). Two touchpoints were selected with the intention of establishing whether student attitudes change over the course of an academic year. The first survey was a paper-based survey, issued in person at the start of the first year of study. This was done in late October / early November, and for many students this was before or around the time of the first submission date for assessed work. Surveys were handed out to approximately 120 students in total from specific classes from each of the participating faculties and the purpose of the survey explained. Students did not have to take part, and were asked to complete consent forms if they were happy to be involved. There were 30 questions in the survey, of which nine relate to personal information (age, gender, course, living arrangements etc), and the remaining relating to student views and experiences regarding plagiarism and academic integrity. Living arrangements and whether the student was working (and how many hours if so) was deemed relevant as this may contribute to stress, which may be a factor in students’ reasons for cheating.

A second survey using the same questions was distributed to the same groups of students using class lists in April 2019. This time, the survey was created using Jisc OnlineSurveys. Links were distributed to 191 students, this number being greater than the first survey point as all students enrolled on the courses used at survey point 1 were sent a survey, whereas the first survey point required students to be present in class on the day the survey was issued. Two reminders were sent out before the closing date. 34 surveys were completed – a significantly lower number than the first survey, which is most likely due to not being a
captive audience during a class and echoes Cohen et al.’s comment about abandonment (Cohen et al., 2018). The full survey can be found in Appendix G: Ethical approval documents.

Findings from the two surveys were collated via spreadsheet and combined to extract quantitative data. In addition, thematic analysis was done on the open-ended questions using Nvivo. Results can be seen in Chapter 4.

2.5.4 Informed consent

Respondents for the first survey were asked to sign a cover sheet giving their consent to take part in the research. This was later removed during the coding process. A tick box on the first page of the survey questions was provided to confirm that the respondent had given informed consent to take part in the survey. On review, this text box was missed by many respondents, and was therefore left unticked. However, it was deemed that since the first page providing student details had been completed, this was most likely an oversight on the student’s part and therefore those results have been included in the analysis. Of the 90 responses received for the first survey, only 38 ticked this box. For the second survey, completion of the online survey itself constituted consent, and this was made explicit in the guidance.

2.6 Methodology for phase 2 – development of ‘Clarify’, a software tool to detect academic misconduct

The next stage of the research was to explore the use of digital forensics and cyber security techniques to see whether they could extract useful information about a document including how it had been put together. The goal was to create a software tool to carry out the analysis, automating it so that assessors require no specialist knowledge, and returning a comprehensive but straightforward report to the assessor.

This phase of the research was carried out through a series of investigations and an iterative approach to software development, with the following stages:

1. Preliminary testing on a known case of contract cheating, based on the theory that digital forensics techniques can be repurposed to analyse a Word document. The hypothesis was that extracting the underlying XML of the document could provide metadata that would be useful in identifying academic misconduct.
2. Investigations to explore whether similar techniques could be applied, theoretically, to copy-and-paste plagiarism.

3. A more detailed analysis of the underlying language (XML) of a Word document to learn more about the artefacts created during document writing.

4. Development of a prototype tool with a working title of ‘Clarify’ to analyse the XML in full and return user-friendly results to the screen. This was tested through a series of iterations as follows:
   a. Testing on researcher-created documents to explore individual aspects of the XML data to develop the appropriate code to draw out these aspects. This was important to ensure the tool was built appropriately and results were accurate, without complicating findings with other aspects of the XML.
   b. Testing on known cases of academic misconduct to confirm the tool functioned as intended on full length, genuine, documents (where ‘genuine’ here refers to the document having been submitted as student work, despite being plagiarised or contracted out).

5. The final part of the software development was to tidy up the code and provide meaningful examples for its demonstration, which form the foundations of a user guide.

During stages 4 and 5 the software was developed and improved in order to create the final functionality. All stages are discussed and explained in detail in the next sections.

This approach is somewhat similar to the iterative enhancement of software development as described by Basil and Turner (1975). In this method, a simple initial implementation of part of the problem forms the starting point, and the software is enhanced through iterative changes until the full system is designed. During each iteration, additional enhancements and additions can be implemented. Basil and Turner suggest the creation of a Product Control List to contain the tasks that need to be completed in order to reach the final desired product. During development of ‘Clarify’, a less formal version of a Product Control List was used by creating a list of ‘To Do’ actions which remained at the top of the tool output screen (webpage) as a constant reminder of actions that needed to be completed. This list was
extended throughout the project, and items removed when completed. In hindsight, a more formal version of this list would have been useful to track changes to the software and to note when actions had been achieved, though this did not hinder development of the tool.

Agile software engineering is also relevant as a methodology here. According to the Agile Software Development Manifesto there are four core values comprising: 1. Individuals and interactions over processes and tools; 2. Working software over comprehensive documentation; 3. Customer collaboration over contract negotiation; 4. Responding to change over following a plan. ‘Clarify’ was not developed with a customer in mind, so values 1 and 3 are of less interest, although it could be said that communication with supervisors and the researcher, as well as an understanding of what the sector was saying in relation to academic misconduct detection was a key driver for the development of the tool. In addition, discussions with the supervisors surrounding the tool development formed a sort of customer collaboration, as did dissemination of the work in progress through conference presentations and research articles. However, value 2 is very relevant. Software outputs and ensuring that the tool worked as expected, with particular emphasis on checking to ensure there were no false positives – as this could be particularly damaging given the purpose of the tool, were the primary aims of the development. Documentation and test cases to provide evidence of its features, along with instructions on how to use the software came later. The fourth value, responding to change, is also key in the software development. As the software was developed, and the possibilities were realised, additional elements were added and existing elements refined, meaning that whilst there was an overarching plan relating to what the aims of the software were, there were no strictly defined plans for development that needed to be rigorously adhered to.

Referring to the proposed activity list that starts this Methodology, in relation to Basil and Turner’s Iterative Enhancement, the following developments were seen: Stage 1 of the above process (preliminary testing on a known case of contract cheating) formed the first simple implementation of the software; a very rough and ready approach to design which established proof of concept, referred to by Basil and Turner as a ‘skeletal implementation’. It was not known at this point whether there would be value in continue the investigations, so unlike traditional software development, no clear objective was defined for the end product. In the second iteration (stage 2) additional testing was done using a similar software process to see if markers for copy and plagiarism detection could also be determined. The third stage of the
method described above focussed more around developing understanding of the OOXML format, and the tests carried out at this stage were used to inform the software development further. The fourth iteration of the software was a formal working model of the tool ‘Clarify’ that provides detailed results to the user. This stage was the lengthiest of the process, as code was written, refined, extended and tested. As usability of the tool increased, so did extensions that added further insight into the document structure. It was very important at this stage not to get distracted by interesting but perhaps not fundamentally useful extensions. For example, the addition of tool tips (comments that appear when hovering over certain elements of the web page on which the results are displayed) was considered and tested for some time, but eventually removed from the final programme as it was not as useful as other features and made the output more confusing to look at. The final iteration of the software was carried out at stage 5 of the method above, where meaningful examples were imported to the index to allow for demonstration of the key features of the software prototype.

A plan of the software development testing is as follows:

- Create skeletal code and test on known case of contract cheating.
- Investigate researcher created documents to draw out specific XML features.
- Investigate student work that has gone through academic misconduct panel and been found guilty.
- Investigate volunteered authentic student work.
- Develop user interface (web page) to summarise findings.
- Test files selected at random to compare automated output with manual analysis to test the reliability of the code.
- Develop a user-guide.

### 2.7 Ethical considerations

All research should be conducted ethically, but this is perhaps even more sensitive in research discussing ethical behaviour and integrity. Oliver (2010) suggests that one of the key issues is to ensure research is based on a firm moral footing, with the intent to improve the human condition. In this particular project, one of the aims was to understand the students’ view of academic integrity and reasons for academic misconduct, as this can help educators to prevent students from cheating behaviours in the first place. The other aim was to create an additional method of detecting plagiarism and contract cheating through a new software tool.
Both these aims contribute to the wider goal of reducing academic misconduct in education, and helping to ensure that students leave with qualifications they have earned. As such, these aims align with Oliver’s suggestion of a firm moral footing.

Ethical consent was obtained for each part of the research, but it was not without its challenges. In particular, the institution where the research was conducted raised concerns about the welfare of students, as well as the implications on fitness to practice if students admitted to cheating behaviours. In addition, and in relation to the testing of the software tool, permission was only granted to analyse the work of students who had either already gone through an academic misconduct panel and been found as having a case to answer, or who were willing to volunteer their work for analysis. Given the consequences of identifying work that looked to be plagiarised or contracted out, it was highly unlikely that any student would submit work for testing if it had been plagiarised or contracted out, and as such the possibility of ‘blind testing’ work was very difficult. There was also the unspoken challenge of bringing an institution into disrepute if research into the extent of academic misconduct is carried out and large numbers of plagiarised submissions are found that passed text-matching inspection and received a passing grade. Published research on such findings may be unwelcome from many institutions.

When gathering information from students who have cheated (or been found guilty of cheating), researchers must be particularly sensitive in their approach. Students in this situation may be embarrassed, ashamed, upset, angry, frustrated, or may be indifferent, but there is understandably a duty of care needed from the institution to which the student belongs, to ensure that no further distress is caused through conversations about their situation. It can also raise uncomfortable challenges for educational establishments. On the one hand is the need to demonstrate that quality is being upheld and academic misconduct is being detected and dealt with appropriately, but on the other hand is the need to protect students and the institution itself. McCabe et al. (2012) commented on the reluctance of educational institutions to take part in academic dishonesty research, suggesting that the reason for this may have been fear of adverse publicity, and that if they discovered they had a problem then they would have to deal with it.

Sutherland-Smith et al. (2019) discuss student experiences of contract cheating allegations through a collaborative project between an Australian University and a UK University. They approached the challenging problem of speaking to students in an ethical and morally
appropriate way by getting student representatives (Student’s Union) and Study Skills staff to conduct focus groups, transcribing them anonymously prior to the research staff (who were academics) reviewing and analysing the results. Not all institutions will find this acceptable though, instead insisting that students who are alleged to have engaged in academic misconduct, who then go on to confess this to be true during a focus group, must be reported through formal processes to ensure that the quality of the institution’s awards is upheld. It is, of course, difficult to capture data from guilty students because of ethical concerns, yet these are the students who can potentially yield the most interesting information, as they are the ones who have found themselves in position of wanting to or feeling that they need to cheat in these ways.

Considerable thought was given to all these issues. During the survey phase, this not only collected data on students’ perceptions of academic integrity, it also asked whether they had engaged in cheating behaviours on any previous courses. It was a deliberate choice to ask about previous courses rather than the current one, partly because the first survey was issued at the start of their undergraduate studies and the students completing the surveys may not yet have submitted any assessment work, and also because asking about previous studies as opposed to current courses mitigated one of the concerns around divulging information that could be harmful to the student. Consideration was also given in relation to whether students could or indeed should be paid for taking part in the research, particularly for the online surveys detailed in the methodology, as an incentive for completing it, and for volunteering work to be assessed by the detection tool during testing. This was decided against though, partly because existing researchers in the field had not routinely adopted the approach of paying students for their time, and partly because there were no specific funds available for this purpose. Other discussions took place with several members of the researcher’s institution including the Data Compliance Officer (to ensure no breach of data protection regulations occurred during the research), Head of Student Casework (to discuss appropriate methods of accessing student work that had been found guilty of academic misconduct) and the Vice President of the Student Union to discuss any potential impact on students.

Turning specifically to the second phase of the research, academic misconduct detection is both an ethical and technical challenge. Current methods of detection rely on the assessor’s knowledge of the student, text-matching software such as Turnitin, and increasingly on authorship tools, which check documents from a single student and across student cohorts for
stylistic differences and similarities. In relation to the development of the software tool ‘Clarify’, thorough consideration was given to whether the process of submitting work to a VLE or to Turnitin acted as consent to take part in forensic review of that work, and whether all work could potentially be checked by the tool as part of the assessment process (in a similar way that Turnitin is used), or whether explicit consent would be required because of the novel nature of the tool. This was an interesting conundrum in many ways, as it is entirely possible to review the document metadata with a little technical knowledge, and one could argue that doing this manually is no different to using an online search engine to search for text, or dropping the text into a text matching tool to do the same (if Turnitin or another text-matching tool is not already being used). There is nothing to stop an assessor doing this of their own volition at any point in a student’s studies. However, because the work formed part of a research study, it was appropriate to gain ethical consent for testing student submissions with the software tool. Ethical approval was granted to use any submissions that had gone through an academic misconduct panel and been found as having a case to answer, without requiring informed consent from students, since the work had already been found guilty of academic misconduct and any further findings would not impact that decision. This student work was analysed in order to help develop the tool and build its efficacy. In order to acquire authentic (i.e. non-plagiarised / non contracted work) from students, informed consent was required, and a request was circulated by means of an online form, distributed by several course leaders, asking for volunteers to submit work for analysis. The form itself detailed the research and submission of a form constituted implicit consent (and this was clearly stated on the form).

In total, ethical approval was sought and granted for four aspects of the research:

- Primary research into student perceptions of academic integrity.
- Analysis of a case of contract cheating using digital forensics techniques and subsequent publication of findings.
- Analysis of work by students that had been referred to an academic misconduct panel and been found ‘guilty’, using the prototype tool, ‘Clarify’.
- Analysis of authentic work volunteered by any student, using the prototype tool, ‘Clarify’.

Ethics forms can be found in the Appendices.
2.8 Summary of Chapter 2

A two phased approach was adopted for this research project. Taking a holistic approach to academic integrity, the first phase focuses on students’ perspectives of academic integrity and the second phase focuses on development of a novel method for detecting academic misconduct. The reasons for redirecting the research into detection were explained as being the result of gaining access to work written by a contracted author and the researcher’s experience in cyber security and digital forensics. A number of methodological approaches were considered, and the rationale for the methods chosen was provided.
Chapter 3: Literature Review

3.1 Introduction

Newton once noted the importance of building on previous knowledge (Newton, 1675). This is true of academic writing: before exploring novel avenues of research, it is first necessary to examine what has gone before and build upon this as a kind of context-setting exercise, highlighting gaps or omissions where they occur. This is a fundamental aspect of research and something students are taught when learning about how to study. As such, students need to be given the right tools and support to help them reference correctly. Whilst the main focus of this thesis is around exploring the application of cyber security techniques to academic misconduct detection, and the development of a novel digital forensics tool for this purpose, an ethical place to start is by taking a more holistic approach to student behaviours. This chapter will therefore begin by reviewing existing literature on student perceptions of academic integrity and academic misconduct, exploring how this information has been obtained and to what extent the information is direct from the student through face-to-face and open-ended questions rather than via closed, quantitative questions. If we can get to grips with student thinking on these issues, and their motivations for engaging in academic misconduct and indeed for not engaging, we may begin to understand how to deal with the problem more effectively. As Gullifer and Tyson (2010) suggest, drawing on psychology theory and thoroughly understanding how an individual views a problem is an essential element for facilitating a change in that person’s behaviour.

Next, there will be a review of existing literature relating to academic misconduct detection tools, identifying the effectiveness of such tools along with any gaps that may exist in this area. The role of technology in the fight against academic misconduct will be discussed, extending into cyber security approaches at HEIs, along with examples of cyber security type approaches to dealing with academic misconduct. The opportunity for these approaches to assist in evidencing academic misconduct will be explored. There will then be a review of literature relating to the application of digital forensics in relation to academic misconduct detection and this will include an overview of the Office Open Extensible Markup Language (OOXML), as this is pertinent to the second phase of the research and an understanding of its current use is essential for remainder of the thesis.
The chapter draws on previously published peer review papers written by the author and co-authored by supervisors, including Johnson and Reddy (2018), Johnson and Davies (2020), Johnson et al., (2022) and Johnson et al., (2023, in print).

3.2 What are Academic Integrity and Academic Misconduct and why do they matter?

Traditionally, education is about the pursuit of knowledge, brought about by a love for learning (Mourad, 1958). This may be in contrast to modern education, which is a means for some students to getting a qualification that supports future employment. Dewey (1916) discusses the value of education at length, including the impact of education on conscious life, development of morals and values, and saying that education is not just “a means to a such a life. Education is such a life”. Dewey comments on education as having two strands – one being the development of specific subject knowledge, and the other being much more important development of critical understanding of the world around us, where school becomes a miniature community which helps to form character and the ability to constantly readjust, which is essential to growth. Rettinger and Bertram Gallant (2022) highlight the problems of focusing too heavily on issuing credentials, and that education can all too often be seen as transactional – “if I complete these tasks, you will award me a degree”. Rettinger and Bertram Gallant (2022) also highlight that the importance of education must be in teaching and learning and that the value of education must be re-established. Academic integrity is, in part, about quality assurance. Assurance that the work submitted by a student and subsequently graded is indeed worthy of that grade, and that the work truly belongs to the student submitting it.
In considering why it is important to care about academic integrity, McCabe et al. (2012, p.3) suggest six reasons as given below:

1. **Integrity is the cornerstone of academia**
2. **Cheating is widespread and on the rise**
3. **The college years are a critical period for ethical development**
4. **College students face significant pressures to cheat**
5. **Students are being taught that cheating is acceptable**
6. **Today’s college students represent tomorrow’s leaders**

McCabe et al. go on to discuss these points in more detail, but these reasons provide a clear and well-argued starting point for discussions around the importance of academic integrity. Foltýnek et al. (2020) suggest that for academic students, plagiarism is ‘detrimental to competence acquisition and assessment’, and that this may lead to professionals who may appear to be qualified in a subject, but who have actually achieved their qualifications based on the work of someone else. They also note the impact on scientific research as impeding scientific process by ‘distorting the mechanisms for tracing and correcting results’, suggesting that plagiarism is not just a problem with students but potentially with academic researchers too. In Garbero (2017), Bretag says that contract cheating in particular is an egregious breach of educational integrity that undermines honesty, trust, responsibility and respect. Furthermore, existing literature suggests that students who engage in unethical behaviour during their studies are more likely to go on to lie to their spouses, employers and customers, to be dishonest when completing expense or insurance claims, and to behave dishonestly at work (Burns, 2022). McCabe et al. (1996) approach this same sentiment by discussing how strong ‘honor codes’ or ethics codes in colleges reduce the likelihood of self-reported unethical behaviour in the workplace.

Newton (2018) looks at past research to determine whether contract cheating is on the rise, and suggests that there is evidence pointing to this, although he recommends that more rigorous and independent studies are carried out to find out the true extent of the problem. Newton’s study results in an estimated 15% of students engaging in contract cheating of some description – that is students who admitted to paying someone else to undertake work for them. Detection rates are significantly lower than this.
Lancaster and Cotarlan (2021) discuss the rise of file sharing websites during the Covid pandemic as a way for students to engage in academic misconduct. Sites such as Chegg (https://www.chegg.com/) allow people to upload a question and have it answered in real time. They found that on average well over 14,000 questions across five STEM subject areas were posted on file sharing website Chegg during 2019/20, an increase on the previous year’s figure of 8357, with a clear spike in the number of questions being asked around March, April and May time. There is also a clear increase in questions posted and answered from April 2020 onwards, suggesting that students are using Chegg ‘as an easy way to breach academic integrity by obtaining outside help’. Their research closes by stating that academic integrity breaches require “a continued reconsideration of teaching and assessment methods” that make it far more difficult for students to gain readily available answers to assessment questions.

It is often suggested that assessment design and the use of ‘real-world’ tasks can help prevent or at least discourage contract cheating (Ellis et al., 2019; Sutherland-Smith & Dawson, 2022). However, Ellis (2019) points out that even if authentic assessment design is used, there are limitations to its ability to prevent contract cheating, though it may act as a deterrent, may make it more difficult or expensive to cheat, and may also make it easier to detect when it does occur. The paper also notes that there is very little evidence to suggest that assessment design impacts the occurrence of contract cheating at all. Another form of assessment that may help to reduce the occurrence of contract cheating and plagiarism is that of oral examination. Research by Schurgers and Qi (2022) into the use of oral examination to support academic integrity found that whilst they did find some evidence of academic misconduct, they found that the examination process helped them to better understand the learning needs of the students as well as students expressing that they preferred this method of assessment. However, Schurgers and Qi note the labour-intensive nature of this form of assessment, and it is clearly not practical if student numbers are high or there is insufficient time allocated for assessments to take place.

The 10 Principles of Academic Integrity for Faculty (Pavela et al., 2017) focuses on virtue, honesty, dishonesty, and ethics, and discusses the impact on assessment type for students, but does not appear to include discussions directly with students. Quality Assurance agencies in both the UK and Australia have drawn up frameworks or guidance for dealing with academic misconduct (TEQSA, 2019; Quality Assurance Agency for Higher Education, 2022), and
there is an excellent summary of the top 42 Academic Integrity articles and book chapters from 1992 to 2012 that was created by Bertram Gallant in 2012, available at the International Center for Academic Integrity, though this is now a decade out of date.

### 3.3 Legal issues

A discussion around academic integrity would not be complete without considering any relevant legal implications. At the time of writing, a number of countries have made essay mills illegal entities and the advertisement of such services an offence. These include Canada, Australia, New Zealand, Ireland, and more recently, England (QAA, 2020). Whilst this does not criminalise contract cheating itself, it makes the provision of work in return for money illegal, and sends a clear message to students that use of essay mills is wrong. It also gives greater clarity for HEIs that see students being targeted with essay mill advertisements, helping them to take action to remove and / or block advertising, and educate students about the risks of using such services. As such laws begin to have an impact on essay mill services, the need for suitably qualified personnel, who can evidence that this has occurred, will be needed, and this becomes relevant to later discussions around detection within this thesis.

There are relevant legal aspects here too from the student’s perspective. For example, any attempt to impersonate a student by logging into their student account (perhaps to complete an online assessment on their behalf, or to submit course work) could be a breach of the Computer Misuse Act (1990, UK), or by similar laws in the country of the activity (e.g. The Criminal Code Act in Australia, 1995, and The Computer Fraud and Abuse Act in the US, 1986). Similarly, sharing login credentials with the intent for someone to use those credentials to carry out activities made illegal under these Acts could also be prosecutable. Students who break into systems to change exam results or to download exam papers from an area that they should not be able to access will likely be in breach of Computer Misuse legislation. There have also been discussions around whether or not contracting to cheat constitutes fraud in the eyes of the law (Draper & Newton, 2017). And yet, despite these legal issues and associated consequences, students will still find themselves in the position of engaging in academic misconduct. Such legal implications for students, as well as for institutions who have a duty of care to their students, require consideration from a moral and ethical perspective too.
Academics across the globe will have varying understanding of the terms academic misconduct, academic integrity, and plagiarism, and this is also true for students. Park (2003) carried out a thorough review of articles that focus on student perspectives, but as will be seen later, there are few direct conversations with students themselves. Adam (2016, p.519) focused on student perspectives, suggesting that students may be confused about what plagiarism is and how they can avoid it, as well as the perceived differences between intentional and unintentional cheating. Adam commented on the lack of research into what students think and understand about referencing and citations.

One of the most common ways to gather students’ perceptions is through surveys. Extensive surveys by McCabe (from 1990 – 2015) found that 36% of students admitted to “paraphrasing / copying a few sentences from Internet sources without footnoting it” (plagiarism.org, 2017; McCabe 2001). Bretag’s review (2014) of numerous surveys notes that the key findings have been that ‘breaches of academic integrity are rife in colleges and universities around the world’, and that little has changed since the early surveys of Bowers (1964) and McCabe.

3.4 Integrity as a writing process

Academic integrity is the cornerstone of this research and there already exists a wealth of information about academic integrity and academic misconduct (Bertram Gallant, 2012; Bretag, 2016; Clarke and Lancaster, 2006; Fishman, 2009; McCabe et al., 2012). Education goes hand in hand with assessment methods, and there is significant work published in this area (Glendinning, 2022; QAA, 2022). It has also been noted that written assessments lend themselves to academic misconduct more than many other assessment formats (Sutherland-Smith and Dawson, 2022), yet at the time of writing, written assessments are still a very popular method of assessment. Ellis et al.’s (2020) five factors of authenticity may help rate the authenticity of assessments, but the authors suggest that authentic assessment does not, in itself, assure academic integrity. The appropriateness of using written assessments as a means of measuring student capability is beyond the scope of this research, but it is worth noting that QAA (2022a, p. 21) identifies the risks of using this format of assessment, saying that whilst they are “valuable, well established and relatively low cost to operate – [they] can provide the greatest risk of cheating”. They suggest that assessors can reduce the risks by including ‘checkpoints’ to review drafts of early work, seminars for discussing early thoughts, including a presentation element, using online testing or workplace attestation.
There are other reasons why written assessments are so popular, including development of the skills that written work encourages. Writing an assessment of any length is a process, and during that process students develop many skills, including researching, reading and summarising, assimilating facts, making judgements and opinions, then leading to creative work of their own. This follows a widely recognised pattern of development as described by Bloom’s Taxonomy (Armstrong, 2010), where in the early stages of learning, students recall facts and basic concepts, then later in the model as their skills develop they begin to apply new information and analyse what they are learning, until such a point where they can produce new or original work (Figure 1). This is what education is all about, particularly at degree level, where students are expected to do more than just recall and restate facts and information, and are expected to be able to make their own judgements and critiques of information they have researched (Biggs and Collis, 1989).

Figure 1: Bloom's Taxonomy from Armstrong (2010)

The process of researching, writing, revising and then submitting a written assessment is therefore a very important part of the learning process, giving students an opportunity to fine tune many of the skills described, and then receive feedback from their tutors on what they have done well and what can be improved in their next piece of work. As such, there is huge value in the writing process, which can help develop higher order thinking skills and self-sufficient learning, alongside deep learning. This writing process is usually not visible at the point of submission. Some assessors will request draft copies of work as part of the submitted
assessment, and some will ask to see drafts along the way. However, the writing of the document is also evidenced within the Word document itself, though this is not usually scrutinised in any way. This is explored in much greater detail in the following chapters.

The writing process itself is as important (if not more important) than the end product, because it is during the writing stage that the student will learn the most. When students copy-and-paste text, they are barely skimming the lower levels of learning, merely scanning text to see if it looks about right before adding it to a document. Similarly, if a student asks someone else to write the work for them, they may not even properly read what has been submitted and therefore will have learnt very little indeed. Both these approaches lead to surface learning at best. Therefore, maintaining integrity in the writing process is a key factor to learning. Yet students’ understanding of academic integrity is not always as good as their tutors may like, as students will often focus on misconduct and the penalties for being caught, rather than the benefits to their own skills and learning of doing the work themselves, with integrity.

### 3.5 Understanding student attitudes – The Student Voice

Given that the focus of the first phase of the research is on hearing what students understand about academic integrity, any form of research that gathers the student voice is useful. As noted in Johnson and Reddy (2018), whilst detection and penalty are very important, a more effective approach would be to tackle the root cause through understanding student attitudes, and the best way to do this is through hearing the student voice in relation to these issues. Understanding more about what students’ views of academic misconduct are enables academics and support staff to provide more timely and appropriate interventions and support. Targeted campaigns at pressure points in the academic calendar, promoting study skills services, support mechanisms and the process for applying for extensions on deadlines would all provide significant benefits. Waltzer and Dahl (2022) comment on the lack of research considering student perceptions of cheating. They approach cheating from a psychological standpoint, posing the question as to why cheating is so prevalent when it is also so widely condemned. In particular, they comment on the importance of understanding student perceptions as their research shows that students appear to be unaware that they have engaged in any kind of academic misconduct.
3.5.1 Student perceptions – a taxonomy of existing research

In the first instance, a review of existing literature in relation to the student voice and student perceptions was undertaken and themes drawn out. The research commenced within a theoretical framework that there is a lack of student voice within the research, with studies concentrating on academic-driven questions posed to students (rather than letting students dictate the direction of questioning) and that responses are often driven by self-reporting students, who are the least likely to engage in academic misconduct. Reviewing current literature will help establish what information exists from students, and how this impacts on the way academics and institutions address the problem of contract cheating.

Sources of information came from research papers, websites and surveys and some significant and relevant studies were identified. A number of large-scale surveys exist which provide a wealth of data. Initially summarised in Johnson and Reddy (2018) and expanded here, these include several studies by Don McCabe, focusing on the United States and Canada (McCabe, 1993, 1999, 2005; McCabe et al., 1996, 1999, 2001, 2012; McCabe & Trevino, 1993, 1997), The Academic Integrity Standards Survey, Australia (Bretag et al., 2014), the Amber project (Tennant et al., 2007), the IPPHEAE Project (Foltýnek and Glendinning, 2015) and GEMS (Global Essay Mills Survey, ENAI, 2017), amongst others.

Table 1 sets out the most significant surveys to have been carried out in this area over the last 30 years. Brief detail is provided around the demographics of the surveys or interviews along with whether these include students. The surveys were checked to see whether they included any open-ended questions as these can provide rich, authentic and honest responses (Cohen et al., 2018) as opposed to being limited to pre-set categories of information, though in many cases these were simply an opportunity to expand on the questions posed (i.e., if you said ‘other’ please elaborate) meaning that they are still fairly directed in nature. Whether or not any face-to-face discussions with students occurred is also noted in the table. These typically provided the most detailed data on students’ perceptions of academic integrity, plagiarism and contract cheating. Within the surveys group, ten major surveys were initially identified, these were added to during later phases of the research resulting in 18 surveys of interest at the time of completion. Each survey is categorised by the number of respondents and whether the respondents were faculty or students, the type of questions (e.g. qualitative, quantitative, open, closed, free text, tick boxes and / or Likert scale) and the purpose of the survey (for example, to establish a student’s understanding of academic integrity, to establish awareness
of policy and penalties, and to find out numbers of students engaging in academic misconduct etc). It became clear from this research that surveys tend to focus on quantitative data, but where qualitative data exists it can be very revealing and informative, although how this data is acted on appears to be less well documented.

What became apparent during this research was that a variety of different foci exist within the literature, including policy, detection, attitudes (both staff and students), forms of plagiarism, language used, academic views, student views, penalties, consistency, ethics and many more, making this one of the most challenges aspects of the research.
Table 1: Summary of a selection of key academic integrity surveys

<table>
<thead>
<tr>
<th>Author</th>
<th>Detail</th>
<th>Includes student responses</th>
<th>Open-ended / free text questions as part of survey / focus group</th>
<th>Face-to-face discussions with students</th>
<th>Focus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowers (1964, cited in McCabe et al. 2001)</td>
<td>5000 students from 99 US colleges and universities</td>
<td>Yes</td>
<td>Yes</td>
<td>Not stated</td>
<td>Prevalence, Behaviours, Context</td>
</tr>
<tr>
<td>Davis et al. (1992)</td>
<td>6000 students</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Prevalence, causes, techniques, HEI response</td>
</tr>
<tr>
<td>McCabe and Trevino (1993)</td>
<td>6096 students, 31 colleges and universities carried out 1990-91</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Perceptions, honour codes</td>
</tr>
<tr>
<td>McCabe (1993)</td>
<td>800 faculty from 16 US colleges and universities</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Perceptions, honour codes</td>
</tr>
<tr>
<td>McCabe, Trevino and Butterfield (1996)</td>
<td>318 alumni of two private liberal art colleges</td>
<td>Yes (alumni)</td>
<td>No</td>
<td>No</td>
<td>Private institutions</td>
</tr>
<tr>
<td>McCabe and Trevino (1997)</td>
<td>1800 students at 9 medium to large universities carried out 1993-94</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Context, honour codes</td>
</tr>
<tr>
<td>McCabe et al. (1999)</td>
<td>31 colleges and universities (repeat of 1990 survey, carried out 1995-96)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Context, honour codes</td>
</tr>
<tr>
<td>Ashworth, Bannister &amp; Thorne (1997)</td>
<td>19 interviews</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Causes, perceptions</td>
</tr>
<tr>
<td>McCabe, Trevino and Butterfield (1999)</td>
<td>Qualitative study, 21 campuses</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Prevalence, behaviours, context</td>
</tr>
<tr>
<td>Brimble &amp; Stevenson-Clarke (2005)</td>
<td>1396 responses (1206 students) from 4 Australian universities</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Prevalence, perceptions</td>
</tr>
<tr>
<td>Lin &amp; Wen (2007)</td>
<td>2068 students from colleges in Taiwan</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Prevalence, techniques, cultural differences</td>
</tr>
<tr>
<td>Christensen-Hughes, McCabe (2006)</td>
<td>14,913 students from 11 Canadian HEIs</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Prevalence, techniques, cultural differences</td>
</tr>
<tr>
<td>Gullifer &amp; Tyson (2010)</td>
<td>Focus groups with 41 students from one Australian University</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Perceptions</td>
</tr>
<tr>
<td>Bretag et al. (2014)</td>
<td>15,000 students, 6 universities</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Behaviour, perceptions, HEI response</td>
</tr>
<tr>
<td>Foltýnek and Glendinning (IPPHEAE, 2015)</td>
<td>3980 students, 27 EU countries, 116 participants in student interviews and focus groups</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Policy, HEI response</td>
</tr>
<tr>
<td>Sutton and Huba (2015)</td>
<td>Focused on cultural differences in approach to academic integrity</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Cultural differences</td>
</tr>
<tr>
<td>GEMS (ENAI, 2017)</td>
<td>Global Essay Mills Survey (2017-)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Cultural differences, perceptions, contract cheating</td>
</tr>
<tr>
<td>Alijurf, Kemp &amp; Williams (2020)</td>
<td>In person discussions, Middle East; interviews with 14 students who had committed academic misconduct; four focus groups with other students</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Private institutions, cultural differences</td>
</tr>
</tbody>
</table>
Of the sources listed, it is perhaps not surprising that very few of the surveys on students were carried out face-to-face. A number of the surveys include open-ended questions, but in many cases these were simply an option to add a response of ‘other’ following a list of types of academic misconduct, reasons for cheating etc. Gathering quantitative data from large-scale surveys is more cost-efficient and time-efficient that individual interviews (Cohen et al., 2018) with a limited number of students, and focus groups are also far less practical. But the risk here is that the information gathered by these surveys is as a result of the questions posed are those of the researcher – i.e., the questions are based on what the researcher thinks the students will want to comment on, rather than allowing a free discourse about academic integrity. Nonetheless, these surveys provide interesting data about why and how students engage in academic misconduct.

The research of Michalska (2014) is particularly interesting as it uses only focus groups and interviews, as opposed to large-scale written or online surveys, from across a number of European countries. Hence the data provided is qualitative, and as such, Michalska notes, most of the participants ‘talked freely’ about their experiences and attitudes towards academic misconduct. Michalska comments on the importance of building a rapport with the students, and that this was easier to accomplish because of the similar ages of the students and researcher. Furthermore, Michalska explores the relationship between nationality and approaches to academic misconduct, as well as drawing up tables of why students may decide to plagiarise or not. She concludes that disparities in opinions may lie in cultural differences and in different approaches and practices at each university. Clearly there are regional differences in students’ perceptions of academic integrity. This is discussed in detail in Bretag’s Handbook of Academic Integrity (2016), which forms a reference work on plagiarism, contract cheating and international perspectives.

Bretag et al. (2014) comment that existing research (at the time of their study) did not fully focus on student perspectives and how this can inform academic integrity policies and practices, something they claim to cover in their study of 2014. Their survey used an online questionnaire which included demographic detail, and understanding of academic integrity.

As can be seen from the Table 1, one of the most prolific researchers in this field is Don McCabe. McCabe and colleagues carried out a variety of surveys from across a number of institutions during the period 1990 – 2015, by the end of this period having surveyed some 70,000 undergraduates. Summarising the data from McCabe’s research between 2002 and
2015, the International Center for Academic Integrity’s website shows that 68% of the 
students surveyed admitted to cheating in some form or another (International Center for 
Academic Integrity, retrieved 2017). McCabe’s surveys used quantitative questions in the 
main, with several open-ended free-text questions at the end (McCabe et al., 1999). This 
appears to be one of the most comprehensive works on student perspectives and includes 
observations from students such as “Provide deterrents to cheating (e.g. harsh penalties)” and 
“Focus on learning, not grades”. Observations like this from students provide us with a little 
insight into student perceptions, though this could go further. In Bretag’s Academic Integrity 
Standards Survey – UNISA (Bretag et al., 2014), comments from students via open-ended 
questions include “I think everyone has a different understanding of what academic 
integrity is and I think that needs to be fixed” (p. 1159), “over my four years at [this university] I have 
always been unsure whether I am satisfactorily meeting the academic integrity policy with 
my work” (p. 1159). Whilst primarily quantitative, this is probably the next-largest scale 
survey available. Carried out in six Australian Universities with around 15,000 students, this 
survey asked students to explain their understanding and awareness of academic policies, 
support available to them, training and timing of training. Bretag reports that further work is 
needed to help students apply their awareness of what academic integrity is to their studies. 
In the South East European Project on Policies for Academic Integrity (SEEPPAI, 2017), 
comments from students include “they [teachers] are lazy to do good exams where students 
can’t cheat …. [teachers] just repeat questions through years”. Comments from students from 
across a variety of surveys and discussion groups suggest that students perceive plagiarism 
(or academic misconduct) less seriously than staff do (Ashworth et al., 1997; Brimble & 
Stevenson-Clarke, 2005; Franklyn-Stokes & Newstead, 1995). Some students also 
demonstrate remorseful behaviour as in Ashworth et al., (1997) – ‘If I knowingly cheated I 
think I would feel guilty, extremely guilty’.

The IPPHEAE study (Foltýnek and Glendinning, 2015) also took the approach of analysing 
the situation across more than one country, in this case focusing on 27 EU countries. Looking 
at the key areas of strategies for countering plagiarism (including awareness, prevention and 
detection); policies and procedures at institutional and national levels, and a review of how 
visible and effective they are; and e-tools in use for detecting plagiarism, this research used 
mixed methods to gather data, including surveys, interviews and case studies. In particular, 
the research included focus groups and interviews with 116 students, and the researchers note 
the ‘very rich data’ which results from this direct contact with students.
The South East European Project on Policies for Academic Integrity (SEEPPAI, 2017) study was a follow up to the IPPHEAE study and looked at just six countries in South Eastern Europe. This focused mainly on exploring perceptions and understanding about policies and procedures relating to academic integrity, with a view to prevent student plagiarism and academic dishonesty. Surveys were carried out through online questionnaires which included a number of open-ended questions, and focus groups using structured interviews were held with students and staff. One finding from the focus groups suggested that cheating is common and that it may go unchallenged in private universities. Also from the open-ended questions, over 28% of responses called for more effective procedures and stricter penalties, applied more consistently. This itself requires more robust and reliable detection methods, particularly in more complex cases of academic misconduct.

The GEMS project (European Network for Academic Integrity, 2017) was launched to gather information about students’ views on essay mills. It was distributed in 22 different languages across the globe. Results are available from the Czech Republic, which received 541 complete surveys to analyse. There are two parts to the survey: the first looks at students’ motivations to study, what they know about essay mills, and whether they have ever used one (or used other outsourced work); and the second part asks students to consider what consequences should be given to other students engaging in contract cheating / use of essay mills. As with many of the surveys listed above, the only option to add free text comes in the form of inviting the student to add a further comment if they wish. The Czech Republic results show that 75% of the respondents are female, and that of all the respondents approximately 14% say that have had someone else complete work for them which they have then submitted as their own. This figure reflects the findings by Newton mentioned earlier (2018).

To summarise then, the most significant research about student perceptions and actions has been carried out by a number of researchers including Bretag, 2014; ICAI, n.d.; McCabe, 1993; McCabe & Trevino, 1993, 1997; McCabe et al., 1996, 1999; QAA, 2017. There are a good number of discussions around the terms ‘cheating’, ‘academic misconduct’ and reasons why students might engage in these behaviours. Discussions directly with students are less prevalent though, especially around why students study and what they hope to gain from their studies. Some of the surveys include open-ended questions (Bretag et al., 2014; McCabe &
Trevino, 1993; Office of Academic Integrity, 2013) but far fewer use focus groups or individual interviews, despite these providing very insightful information.

This gap in research led to the first phase of the research study, that of a primary survey directly for students, to gather their perceptions of academic integrity by asking some open-ended questions to establish if the researcher’s institution mirrored findings from existing research and to allow students to provide more detailed answers than closed questions would permit.

3.6 Detecting academic misconduct

Given the arguably impossible task of preventing academic misconduct, the research turned to detection of academic misconduct and how this is currently done. The importance of having detection mechanisms in place is not in question. As Lancaster (2022b) points out, having detection methods in place is important, even if only intended as a measure to disincentivise students from cheating in case they get caught. Existing detection focuses almost exclusively on the visible content of the document itself, through examination of various aspects of the text and in some cases the citations and images / equations. There are a number of tools that are widely used to support this activity. One of the most prolific means of plagiarism detection is through the use of text-matching software (TMS) tools, with examples being Turnititin (https://www.turnitin.com), PlagScan (https://www.plagscan.com) and Urkund (https://www.urkund.com). Text-matching is done by scanning archives (whether online or in repositories) for passages of text that appear in student submissions and returning a ‘score’ or percentage match to the assessor. However, by changing one or two words, or switching phrases around, students can outwit the mechanisms used by these tools and achieve low scores, which then in turn fail to alert assessors that potential plagiarism has occurred. New approaches to detecting academic misconduct are important because students are now far more aware that simply copying and pasting work will be easily detected using text-matching software.

One of the major benefits of text-matching tools is the ease with which they can highlight or flag documents of interest, something that would be very challenging to do at scale for individual assessors. Although as Rogerson (2017) notes, whilst detection using technology is reasonably effective when text has been taken from Internet sources, there is a danger of over-reliance on the results as the sole means of detection. Furthermore, documents
Understanding academic integrity and plagiarism in the digital age

submitted in portable document format (PDF) further complicate text-matching, with the text needing to be extracted before it can be checked. Rogerson begins to explore stylistic features of submissions, noting the importance of observing language use and consistency such as shifts in font size, gaps in the document, changes in grammar, tense and spelling. She finishes by saying that it is much about educating ourselves about the ‘patterns, markers and clues’ and that knowledge of the students is very helpful in regards to detection. Of course, this knowledge of students becomes much more difficult when working at scale.

The number of users of Turnitin demonstrates the importance of detection tools to facilitate swift and automated checking of documents, flagging works which show high similarity to other sources, thus enabling assessors to quickly identify which papers should be scrutinised more closely. Turnitin (n.d.) themselves claim that it is used in 98% of UK Universities, and 6 out of the top 10 FE colleges, with an estimated one billion student paper submissions in 2018 according to Think Insights (2023). Of course, as with all tools, they can provide only a flag and results must always be scrutinised by the assessor before claims of academic misconduct are levied.

For cut-and-paste plagiarism, text-matching software is the predominant method of detection used. However, Turnitin has now rolled out authorship investigation tools, which help in the detection of contract cheating using stylometry and other semantics to establish ownership of a document. Turnitin claim to give ‘data-backed insight into whether students are doing their own work’ by helping instructors ‘verify their concerns and easily escalate cases for investigation’ (Turnitin.com, 2021). Turnitin states that this is done through document forensics and metadata, and by using forensic linguistic analysis and Natural Language Processing, and perhaps addresses Dawson’s comment that “Custom written work is increasingly difficult for educators to detect and turn into real evidence” (Turnitin, 2021). Turnitin offers a variety of other tools, including ‘Originality’, which can analyse a submission by one student and compared it to other submissions by the same student noting any anomalies in style. They can also detect work from several different students that bear the same authorship fingerprint (Foltýnek et al., 2019), suggesting that the work has been done by the same person. All these tools rely on the textual content of the document (and some limited use of document metadata), and may include examination of grammar, sentence construction and language, amongst other things. The limitations of all these methods are
already documented by Dawson (Turnitin, 2021), Foltýnek et al., (2019); Foltýnek et al., (2020), and Rogerson, (2017).

3.6.1 Intrinsic and extrinsic methods of plagiarism detection

There are two papers by Foltýnek which are particularly useful in relation to plagiarism detection tools. These are Foltýnek et al. (2020) – ‘Testing of support tools for plagiarism detection’, which uses sample documents and tests them against 15 online plagiarism checkers to establish their effectiveness; and Foltýnek et al. (2019) – ‘Academic plagiarism detection: a systematic literature review’, which is a thorough review of existing research into plagiarism detection tools, exploring different forms of plagiarism, and explaining in detail the contrasting methods of extrinsic and intrinsic plagiarism detection methods. The 2019 paper, in particular, provides a critical evaluation of the “capabilities of computational methods to detect plagiarism in academic documents”, in which Foltýnek et al. note there is a prolific amount of research, but which changes rapidly according to new techniques and software. Their work also aims to identify any gaps in the research. It provides a thorough analysis of plagiarism detection literature from 2013 to 2018, plus any seminal works before this period, and the authors claim that the research includes significantly more papers than other reviews have previously. In accordance with Fishman (2009) they define academic plagiarism as “the use of ideas, content, or structures without appropriately acknowledging the source to benefit in a setting where originality is expected” (p. 7).

In extrinsic plagiarism detection, the document is compared to existing work in order to evidence plagiarism, whereas intrinsic plagiarism detection analyses the input document using ‘stylometry to examine linguistic features of the document’. This is of particular interest to this research, though Foltýnek et al’s suggestion that stylometry is the only technical option for discovering potential ghostwriting (or contract cheating) is questionable, since there are other intrinsic detection methods that can be used (though they currently are not, and this is something that this thesis explores in detail in Chapter 5). Foltýnek, et al. (2020) note that a crucial presumption of the intrinsic approach is that it is possible to identify authors through their different writing styles, so differing styles of writing can suggest that one or more authors wrote the paper, and furthermore that each author has their own distinct fingerprint and it is possible to work out who wrote the different stylistic sections using this knowledge. The main limitation of the methods described by both papers by Foltýnek et al. (2019 and 2020) is that they focus exclusively on the textual content of the
document, so whilst analysis is thorough and detailed, it does not include analysis of what is going on behind the scenes. As noted in their research, there is a consensus in the literature that the inability of plagiarism detection systems to identify obfuscated plagiarism is currently their most severe limitation.

Chowdhury and Bhattacharyya (2018) define two forms of plagiarism – that of ‘textual plagiarism’, which re-uses text-based phrases taken word-for-word without direct quotation and includes copy-and-paste plagiarism, paraphrasing, and re-use of ideas claimed as original; and, secondly, that of source code plagiarism, which re-uses code written by others (such as computer code or computer programming). As in Foltýnek et al. (2019), Chowdhury and Bhattacharyya (2018) describe two main types of plagiarism detection as being intrinsic and extrinsic, where the intrinsic approach analyses stylometrics, searching for various differing writing and authorship styles within a single document as evidence that two or more authors were involved in writing the contents (whether knowingly or otherwise), and the extrinsic approach focuses primarily on attempting to match the contents of the document in question with external resources in digital repositories. It is worth restating that both intrinsic and extrinsic detection methods as described here focus on the text content of the document.

Figure 2 outlines the different forms of plagiarism according to Foltýnek et al. (2019), highlighting the detection methods that can be used for each type of plagiarism (including ghostwriting or contract cheating). Foltýnek et al. suggest that the only technical option for discovering potential ghostwriting (at the time of writing) is to compare stylometric features within a suspect document with those of previous work known to be done by the alleged author.

Of the methods mentioned below, Foltýnek et al. (2019) explain what information the various detection features provide. For example, lexical detection (focusing on words and language) includes analysis of n-grams (a sequence of N words, such as ‘thank you’, which is a 2-gram), which help to determine where styles change within a single document or across documents submitted by the same student; and Vector Space Models, where text is represented through algebraic modelling, often used for information filtering, indexing and retrieval – this method can be used to match one document to another in terms of style and language. Stylometry focuses on linguistic features of the text that indicate a change in the writing style. Non-textual Feature Analysis looks at content that is not text, such as equations, figures and tables. Semantic Graph Analysis represents text in graph form, whilst LSA (latent
semantic analysis) and ESA (explicit semantic analysis) again examine the words used, along with typical word groupings and synonyms to search for similar texts. Finally, the use of machine learning can help combine the best-performing detection methods but again this focuses entirely on the text that is within the document being reviewed and other external sources of text.

![Diagram of plagiarism forms and detection methods]

Figure 2: Suitability of detection methods for identifying certain forms of plagiarism, Fig 2 in Foltýnek et al., 2019

3.6.2 Existing tools and techniques used in plagiarism and contract cheating detection

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 albeit not 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”, and Crockett & Best (2020) summarise that initial identification is often dependent on the awareness of individual assessors: 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. However, teaching large cohorts can make this knowledge of the students’ abilities very difficult. Other techniques include the use of online search engines, 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 or other search engines to see if a match can be found and a viva voce with the student to establish the extent of their knowledge and whether that is consistent with the work submitted (QAA, 2022a). These latter techniques are time consuming for the assessor and cannot be performed in bulk on numerous assessments, and methods that rely upon assessor observations can be difficult in large classes or where assessors do not know the students they are assessing.

Chowdhury and Bhattacharyya (2018) discuss various methods of plagiarism detection, classifying eleven distinct types of detection. All eleven methods discussed focus on text, whether that be through sentence structure, grammar or stylometry. None of the methods includes any technical investigation of the document’s forensic artefacts. Foltýnek et al. (2019) provide a list of literature reviews that focus on plagiarism detection tools. They also note that intrinsic detection methods exhibit a number of shortcomings including a propensity for errors in intentionally collaboratively written documents, difficulties in detecting plagiarised work that has been paraphrased and a general unreliability of the findings at present. Finally, they note the challenge in identifying a potential source document meaning that whilst the tools can flag uncertainty about the document’s origin, extrinsic methods will still need to be used to locate the original source. QAA (2022a) note that the interrogation of metadata around document creation is useful and should look ‘realistic’ but they do not expand on what sort of metadata they are referring to nor how to explore it in any detail.

Chowdhury and Bhattacharyya (2018) provide a list of 31 plagiarism detection tools that were available at the time of writing their research. They explored whether each tool used extrinsic or intrinsic detection methods, and compiled the information into a table. In 2020, Foltýnek et al. also carried out a review of plagiarism detection tools. Compiling this additional list of tools with the list created by Chowdhury and Bhattacharyya (2018) results in a list of 37 unique tools. Of these 37 tools, 35 use extrinsic detection methods only, one uses intrinsic methods only, four use both extrinsic and intrinsic and one from Foltýnek et al.’s list could not be located. None of the tools looked inside the document XML to carry out a more detailed forensic analysis. XML refers to Extensible Markup Language, which is a markup language designed to store information about a document so that it can be transported over the Internet and reconstructed in the way it was originally intendent, a bit like a set of instructions. A full list of the tools reviewed by Chowdhury and Bhattacharyya (2018) and
Foltýnek et al. (2019) is provided in Appendix A: Plagiarism tools (intrinsic and extrinsic), and a summary is provided here in Figure 3:

![Figure 3: Plagiarism Detection Tools – Extrinsic and Intrinsic methods](image)

### 3.6.3 Using stylometrics to detect contract cheating

Contract cheating brings with it additional challenges. Through analysing previous work by the submitting student, it should be possible to compare work and look for similarities in style: vocabulary, sentence structure and so on. However, this is not an exact science as there may be no previous work by the student to compare with, and the student’s style may evolve over time. The conference paper by Juola (2017), seeks to formalise this approach in order to show that two papers have been written by different people.

The linguistic approach is that authors have very unique ways of expressing the same thing, and that their word choice, sentence structure and narrative can be analysed and formalised. If you were to ask an assessor whether they think they have ever felt that a student submission did not ‘feel’ as though it belonged to the student in question, you would probably get a positive response. However, this can be tricky to prove. Other features that should be consistent by a single author include European and US versions of works (e.g. ‘pavement’ versus ‘sidewalk’; ‘rubbish’ versus ‘garbage’ or ‘trash’); varying use of punctuation marks (e.g., different types of hyphenation: en-dash, em-dash); commas before ‘and’, and so on. Juola (2017) describes how two people were asked to write down how to make a ‘PB&J’ (peanut butter and jelly/jam) sandwich, one person was an 11-year-old
female, the other was a 29-year-old male with an advanced degree. The style of writing is completely different, though the task is the same. Juola goes on to discuss frequency distribution (where two papers can be compared for their use of specific words as a way to indicate who wrote which paper). Juola also notes several case studies, where it has been possible to apply a system of analysis which makes it highly likely that literary works found under other authors’ names (or with no attribution at all) can be attributed to specific authors. Indeed, this type of ‘signature’ is used in forensic analysis of computer malware to facilitate attribution to a specific group or nation state (Langweg, 2012). Juola (2017) has developed a system to assist in finding the most likely candidate for authorship using JGAAP (Java Graphical Authorship Attribution Program). This can either analyse a student’s submission against previous submissions by the same student or analyse a student’s submission against the rest of the cohort to search for similarity of writing style with another student in the group. Juola notes that as with many methods of this type, there is a risk of false positives. Notably though, as with intrinsic and extrinsic methods described earlier, Juola’s system only uses the content of the document itself (i.e., the text content) and like the tools summarised in Figure 2 and Appendix A: Plagiarism tools (intrinsic and extrinsic), does not review the XML data behind the document, which can yield rich data specific to the document in question and is discussed in 0.

3.6.4 XML analysis

In Figure 3, a list of tools was summarised as to whether they used intrinsic or extrinsic methods of detection (or both). However, as noted above, not one of these 37 plagiarism / contract cheating tools analyses the underlying XML and more hidden document forensic artefacts within the student submission. There is, however, some research into the use of XML for plagiarism detection, but this focuses almost exclusively on extracting blocks of text for stylistic analysis through averaged word frequency classes (Meyer zu Eissen & Stein, 2006), n-gram analysis (Bensalem et al., 2014) or citation analysis (Gipp et al., 2014), or as a tool for verifying other methods of checking or for providing extraneous information about a corpus of work (Basile et al., 2009; Zouaoui & Rezeg, 2022). The exception to this is the author’s own previously published work, carried out during her PhD studies (Johnson & Davies, 2020a; Johnson & Davies, 2020b; Johnson et al., 2022).

A deeper exploration of XML extends analysis into the Revision Save Identifier (RSID). RSIDs are hexadecimal values or codes that are added to a document as it is built and edited
to facilitate document sharing and revision history tracking. Whilst not seen in the final document, the information exists in the background and can provide useful metadata about the document. This feature is explained in full in Chapter 5. Research does exist into XML RSID analysis, but this is far more limited and tends to focus on gathering evidence for forensic investigations (Jeong & Lee, 2017), document history tracking techniques to identify forensic similarities between two documents (Joun, Han et al., 2018), and establishing connections between corporate documents allegedly from different organisations with a focus on providing evidence for criminal investigations (Joun, Chung, Park & Lee, 2018).

Returning then to the table devised by Foltýnek et al. (2020), the author proposes that this could be further enhanced to include a different type of intrinsic plagiarism detection: that of XML analysis, applying XML investigations to the detection of plagiarism, and that it could be redesigned to look like Figure 4:

3.7 Cyber security and its relevance to academic integrity

Whilst linking academic integrity to cyber security may not seem like the most natural thing to do, there are many synergies that make the link worthwhile. Johnson et al., (2023, in print) discuss this at length in their work, and links to cyber security are also considered by Lancaster (2022b), Dawson (2020) and Murdoch (2019), and cyber security techniques in the form of digital forensics are discussed in research by Klopper (2009), and Didriksen (2014).
Initially an area focused on by a single department (or even a single person), organisations now appreciate the need to take a holistic approach to cyber security. No longer can the functions of the cyber security team be handled by a single entity, rather, there must be a range of approaches that form a layered approach to an organisation’s security. In its most basic form, cyber security focuses on the CIA triad – Confidentiality, Integrity and Availability of data and systems. This echoes many of the needs of academic integrity and Dawson (2020) refers to this framework, noting that the I (integrity) focuses on ensuring that ‘data are not tampered with or deleted by unauthorised parties’. Johnson et al. (2022) take this a step further by adding that the CIA triad elements can be related to academic integrity through the following: Confidentiality – ensuring that intellectual property, exam data, student records etc are protected; Integrity – ensuring that work submitted was written or created by the person who submitted it; and Availability – ensuring that students and staff can access learning resources and the Virtual Learning Environment (VLE) when needed, but that unauthorised persons cannot. Dawson (2020) also references a number of other terms that are often discussed in conjunction with the CIA triad including ‘authentication’, ‘access-control’ and ‘non-repudiation’, all of which are very relevant in academic integrity assurance.

Agencies that support HEIs such as Jisc (UK) and Educause (US) are increasingly recognising the need for support to be provided to improve defence against any form of cyber attack, providing guidance and frameworks to mitigate cyber attacks. The increasing use of online systems for examination purposes has brought the need for proctoring to the fore, with these providers including some fairly sophisticated techniques such as eye-tracking software, multiple face detection and detecting activity outside of the recognised software to ensure that students are not cheating during exams. Johnson et al. (2022) note that these advances in technology that make it harder to detect academic misconduct necessitate the development of a more forensic approach to detection.

Furthermore, cyber security is a continual challenge that HEIs will face, with an ever-changing attack surface and constantly evolving attack vectors. This mirrors academic misconduct, which takes a variety of forms across every discipline, and with students becoming ever more inventive in how they attempt to obfuscate their actions. The transdisciplinary nature of plagiarism has been discussed in a previous chapter, and learning from one field of study or concern can be very useful, resulting in cross-pollination of
knowledge, skills and techniques that can benefit beyond their originally intended use or application.

Johnson et al. (2022) discuss in detail the use of cyber security techniques as the next defence against academic misconduct. Their suggestions include greater use of learner analytics by tracking student logins, as introduced by Murdoch and House (2019). For example, monitoring the IP address of logins can flag ‘risky sign in’, where a user has made multiple failed login attempts with an incorrect password, or has failed multifactor authentication. It can also flag ‘impossible travel’, which is where logins have been made to the same account from different geographical locations in a time span that would be impossible (e.g., a login from a UK based IP address followed by a login from the same user two hours later in a location that is greater than two hours away from the UK). They explore possibilities around big data and data analytics. They also discuss the use of digital forensics in some detail, as a way of unpacking a student submission to its component parts (those of XML and RSID revision identifiers) as a way to examine the submission in far greater detail than just looking at the textual content itself. They explore how a security information and event management system could provide evidence, flagging unusual activity, blocking and / or tracking known IP addresses and highlighting the risky sign in and impossible travel mentioned above.

Finally, they suggest a Defence in Depth approach to academic misconduct, in the same manner that is applied to the cyber security of an organisation, using a number of layered defences against academic misconduct, rather than relying on a single tool. As such, they suggest areas of control that are used in cyber security which could be repurposed and applied to academic integrity:
Table 2: A Defence in Depth approach to academic misconduct, from Johnson et al. (2022)

<table>
<thead>
<tr>
<th>Control</th>
<th>Existing Measures</th>
<th>Additional Cyber Security Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical</td>
<td>Checking ID cards at exams, preventing physical access to course materials</td>
<td>Blocking USB ports to prevent use of keylogging devices; using biometrics for identity validation</td>
</tr>
<tr>
<td>Technical controls</td>
<td>Secure assessment testing environments; multifactor authentication</td>
<td>Blocking unwanted incoming IP addresses (e.g. known essay mill IPs); setting user permissions to avoid access to exam areas, gradebook etc); proctoring tools; behavioural analysis for highlighting anomalous behaviour online</td>
</tr>
<tr>
<td>Administrative controls</td>
<td>Policy and procedures relating to academic integrity and technology (e.g., copyright infringements; acceptable use policies; academic misconduct policies</td>
<td>Risk Assessment, incident response planning; threat intelligence; tabletop exercise; training</td>
</tr>
</tbody>
</table>

### 3.8 A forensic approach to academic misconduct detection

We have seen that existing detection methods focus on the text content of the student submission, but as technology advances and students become more innovative at hiding their attempts to engage in academic misconduct, we need to heed the words of McCabe et al. (2012, p. 13) who note that ‘cheating has always been and will always be with us’. They suggest that we cannot be naïve in our approach to academic integrity. Whilst academics may wish to put sufficient measures in place to prevent academic misconduct, this is a lofty ideal, and detection forms a necessary part of investigating breaches in academic integrity. Given
the limitations of current detection methods and extrinsic and intrinsic approaches, looking for a more forensic investigative approach seems worthwhile.

Weber-Wulff (2016) discusses technological solutions for the determination of plagiarism, and the difficulties these bring. Weber-Wulff notes that there are many, varying definitions of plagiarism, and as such, no one tool can definitively determine that plagiarism has occurred. She suggests that these tools may identify the use of words from other sources, but that human intervention will always be necessary to confirm plagiarism has occurred.

Historically, students have been quite ingenious when trying to outwit text-matching software. In “Turnitoff” (Heather, 2010), Heather discusses plagiarism detection software and talks about a loophole that allows a document to bypass plagiarism checks. He identifies the process of detection through three distinct stages:

1. extract the text from the document (usually a PDF or Microsoft Word document)
2. search for the text online and (possibly) in a customized database
3. quantify the extent to which extracted text could be found in other sources, and display results.

Heather notes that abstracting the text can be the most problematic. He identified that it was possible to stop text from being extracted from the document by modifying the character map of the document, rearranging the glyphs (so that an x would appear as an ‘e’ to the reader but would be interpreted by the plagiarism detection software as an ‘x’), or by replacing all the curves in a document to their Bezier equivalents, thus effectively turning the document into a series of squiggles, which just happen to look like letters to the reader. This is discussed in a Reddit post from 2015, where AtomikRadio (2015) posts a question about a student who received a 3% similarity report in Turnitin, despite the fact that the student submitted a verbatim copy of a report from the previous semester, submitted by a different student (Reddit post, 2015). AtomikRadio noted that there were no white-text spaces, no macros and no paraphrasing / synonym replacement (methods sometimes deployed by students in an attempt to obfuscate their actions). Suggestions in the comments are that there may be a remapping of characters in the PDF version of the document (as per Heather’s paper), or that the file was submitted as an image. Turnitin is now much better at identifying such techniques to avoid detection, but they are interesting nonetheless.
The importance of the data behind word documents is becoming increasingly interesting. Combining deep learning, natural language processing and AI techniques, Docugami (‘The Document Engineering company’) can examine documents, categorise them and ‘identify common and unique elements’ (Costine, 2020) – “You can take all the data that lives inside a PDF that isn’t recorded anywhere else, no one knows how to find, and no one knows how to aggregate, and all of a sudden you can run a query on this unstructured data as easily as if it was a spreadsheet.” Docugami themselves call this ‘Document Engineering’ claiming to be able to create a hierarchical data representation of each document in its entirety using AI, XML and cloud technologies. Docugami are planning to use the information behind the docx and PDF formats for data mining and data aggregation (in much the same way that Google does for the Internet). It is interesting to note that in May 2020, Grammarly announced investment in Docugami (Bishop, 2020), and what looks like a potentially great tool for combating plagiarism and contract cheating now appears to be in collaboration with one of the very organisations that perpetuates this problem.

Meyer zu Eissen and Stein (2006) mention cryptographic hash functions in their research on intrinsic plagiarism detection, where text chunks are ‘hashed’ to find a specific value or digital fingerprint which can then be matched to digital fingerprints in other texts. However, the hash value will only match an exact copy, so even a changed space or hyphen or word will render the hash values completely different. These issues severely limit the effectiveness of their use. Their research suggests the use of stylometry, in particular the introduction of the Averaged Word Frequency Class, and uses XML documents to carry out their experimental analysis. This work is an early example of the use of intrinsic analysis (i.e., using only information available inside the document itself, rather than comparing the document to other external sources).

3.8.1 Word metadata

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 (or metadata) 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 (Figure 5).
However, such information is not always reliable, and can be changed by the author with a basic knowledge of how Word stores this metadata.

### 3.8.2 Language and paraphrasing as detection evasion mechanisms

One area that text-matching and stylometrics fail to address is that of cross-lingual plagiarism. This is where students will take work written in one language, and translate it into another language, using a tool such as Google Translate (Prentice and Kinden, 2018). This may be done because the student’s native language is not the language in which the work needs to be submitted (as may be the case for international students, for example), or as a deliberate attempt to evade plagiarism detection by subtly (or in some examples, less than subtly) altering the structure of the text, thus making it undetectable by text-matching software. Another technique that students may use to evade detection is by using online paraphrasing tools, as discussed by Prentice and Kinden (2018). This plagiarism may be easily detected by the assessor if the language or terminology no longer fits with what the assessor expects and is dependent on the quality of the paraphrasing or translation tool. Ultimately though, the existing tools for plagiarism detection are only examining the text that is submitted, and are not looking behind the scenes for any digital forensics evidence of authenticity.
3.8.3 Examples of digital forensics use in other fields

During the literature review it was possible to locate various articles that discuss forensic techniques used in a variety of circumstances, but there was very little research into its application for detecting academic misconduct. Despite this, the techniques in use show potential to be used in this arena, and as such the next section reviews the information already available in this field. According to Interpol, the goal of digital forensics is to ‘extract data from the electronic evidence, process it into actionable intelligence and present the findings’ (Interpol, 2022). Typically, these methods are used in criminal investigations, but they can be useful in a number of other settings.

3.8.4 The Daubert Standard and linguistics

This section will cover an overview of digital forensics tools and how they have previously been used in cases of intellectual property, criminal investigations and more.

The Daubert standard is a rule of evidence regarding the admissibility of expert witness testimony, based on five key points relating to the procedures, standards and protocols used in the investigation. A requirement of a forensic report, whether for physical (wet lab) or digital (dry lab) evidence, is that the results can be reproduced by an independent third party – inferring that the methods are robust and reliable enough to provide an unequivocal conclusion (Garrie, 2014). This is very important in legal proceedings as it is often the evidence which is used to prove guilt or innocence. However, in the field of linguistics, there are few such standards and recognised procedures, making it challenging to meet Daubert standards, and thus bringing the reliability of the evidence presented into question. The very nature of the subjectivity of language, as discussed earlier, makes reproducing the same results very unlikely.

Coulthard (2013) discusses the difference between lexical words (content, made up of nouns, verbs and adjectives) and grammatical words (which bind the lexical words together like glue, made up of articles, pronouns and conjunctions). There are many more lexical words than grammatical ones, and it is therefore the lexical words that are used to help determine linguistic authorship attribution (Coulthard, 2013). Coulthard explains how single word choices are not especially helpful in determining authorship, but that co-selections of words or phrases is much more useful, and different authors will favour different co-selections. Authorship can be further supported through misspelled words, or common typographical
errors – for example ‘teh’ instead of ‘the’. Some of these errors are more helpful in
determining authorship than others. In terms of student submissions, a D-grade student who
always spells ‘retorical’ without the ‘h’, then submits an outstanding piece of work where
‘rhetorical’ is spelled correctly throughout could rightly raise suspicions of authorship.
Coulthard notes the obvious limitation of this – which is that of the spelling and grammar
checkers now provided with almost all word processing software, making obvious
typographical and grammatical errors far less common. Coulthard was able to draw
conclusions about the authorship of an email using these and other similar linguistic
techniques, but lacked the robustness that is required by the Daubert standard, in that some of
the key linguistic features which were crucial to the expert view were not included as they
were discussion points based on rarity of specific linguistic items – and as such the results
were unlikely to be replicable by a third party. However, Coulthard further developed his
analysis by adding a comparison of the similarity and diversity between works supposedly
attributed to the authors in dispute using a simple Jaccard similarity coefficient which makes
it possible to attribute authorship more reliably by measuring the number words two
documents have in common.

Belvisi, Muhammad and Alonso-Fernandez (2020) carried out some interesting authorship
analysis using digital forensics techniques within microblogging texts (e.g., Twitter). In their
research, forensic techniques using n-grams and stylometric features were employed to see
whether it is possible to identify the author of a message with any degree of accuracy. Whilst
a small-scale study, the initial results showed some promise, yielding a classification
accuracy of between 92% and 98.5%. As with many of the other applications of forensics
though, this focused purely on the content on the message, though the authors noted the
importance of detecting authorship in cybercrime, education and fraud detection – all of
which are relevant to academic misconduct.

3.8.5 OOXML overview

This section provides a brief overview of the Office Open XML (OOXML) format, which is
the format used within a Word document to store information about the document itself
(metadata). In Chapter 5, a more technical explanation of OOXML will be provided as this
forms the basis for the investigation and testing phase of this research. Microsoft Word from
2007, along with other Microsoft packages such as PowerPoint and Excel, use the ‘Office
Open XML Format’ (OOXML) format. The format can be easily identified as all files are
saved by default with an ‘x’ appended to the file extension – e.g., file1.docx, file2.xlsx. A Word document in OOXML format 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. Alongside these are forensic artefacts that cannot easily be changed to obfuscate misconduct-related behaviours, making them very useful not just in detecting copy-and-paste plagiarism but in detecting contract cheating and potentially documents created by artificial intelligence too. Analysis at this level, beyond the textual content of the document, and exploring the editing metadata in particular, is something completely novel in terms of plagiarism and contract cheating detection, though it has been discussed in other contexts.

3.8.6 Application of OOXML in Copyright disputes

Fu et al. (2011) proposed a method for investigating and tracking the source of illegally copied documents using OOXML in a forensic capacity. They also presented methods to help determine the time and creator information for the document, which could be used to establish true ownership of a document and could be used in copyright disputes. The research focuses on the analysis of OOXML and examines the two questions of whether two documents (one suspicious) come from the same source, and who is the real copyright holder (i.e. which document is the original, and which one has been copied). Fu et al. (2011) noted the significance of the Revision Identifier (RI), which is a unique value, randomly generated when a document using an OOXML format is edited. Their experiments show that the w:rsidR value will not change once created, provided that at least one printable character remains unchanged. This then means that identifying identical w:rsidR values across two documents indicates that the two documents must have come from the same source. Fu et al. also note the significance of two files that are packaged in the OOXML format. Firstly, the document.xml file, which contains the main body of the document, and secondly the settings.xml file. Both contain RI values, with the settings.xml file recording revisions of the document, retaining RI values linked to deleted printable text as well as current RI values. They state that it is possible to determine that the copy and the original document are from the same source using this process.
Fu et al. (2011) also discuss the time stamping of OOXML files, noting that when extracted, the Date Modified shows when any part of the document was modified. However, they note that it is possible to edit this information, so a more detailed analysis is required. This can be found in the core.xml file. The creation date of the document can be found in the dcterms:created element, whilst the modified date can be found in the dcterms:modified element. The core.xml file also includes information about the creator (though this may depend on whether the computer system used includes user information) along with the number of times the document was revised.

This refers to two methods – one where an original document is taken, content deleted and changed, but the RI proves the link to the original document, and one by creating a new document and copying information across; providing at least one character remains whose original formats have not been cleared, there will be identical RIs present.

Fu et al. (2011) created a tool to perform this analysis, but the tool is no longer accessible by the link provided in their research (Figure 6).

Figure 6: Digital Forensics OOXML analysis tool, Fu et al. (2011)
3.8.7 Application of OOXML in other contexts

Didriksen (2014) suggest the use of OOXML as a way of connecting the actions performed, e.g., editing the document, “to a specific physical person or several people” when carrying out a digital forensics investigation, as this would allow investigators to attribute certain actions to specific users. Langweg (2012) used XML to establish ownership of documents relating to a terror attack.

Research by Xiang et al., (2016) discusses the use of these techniques for hiding data within a Word document (steganography). Such methods can be used in criminal investigations, but no evidence was found to suggest that they are ever used in establishing that academic misconduct has occurred. Xiang et al., (2016) suggest the use of extensible markup language (XML) as a cover medium to ‘transmit secret information by offenders’ and discuss tools and techniques to detect deliberately hidden information. Similarly, Castiglione et al., (2011) explain how data can be hidden in a single file using the OOXML format. Jeong and Lee (2017) discuss the use of digital forensics techniques in relation to version history to suggest theft of intellectual property through establishing a document content’s creation, modification, deletion, and copy information, though this requires access to the file system of the computer where the document was created, which is unlikely to be the case in academic submissions. Zhangjie et al. (2015) discuss a tool to prevent covert communication using OOXML forensic techniques.

There are some additional tools that are available for forensic XML analysis, though their suggested uses are not in relation to academic misconduct detection. These tools are detailed by Didriksen (2014) and include the following:

- read_open_xml.pl is a Perl script freely available on GitHub that can read a document written in OOXML format and display the metadata information that is contained in it.

- A Google Chrome extension called OOXML Google Extension that allows users to view, edit, analyse and compare contents of OOXML documents, and is designed for software developers who need to work on OOXML documents, without the need for manual extraction of the component files. This tool can be useful in academic integrity cases involving collusion to check the underlying similarity of two documents.
• DOCXRevisions, an unpublished tool created by Langweg (2012), and DSO Tool: Detector for the source of OOXML file (designed by Fu et al., discussed above), which can be used to detect collusion.

Other tools include Autopsy, which is a well-established and reputable digital forensics programme, often used as a verification tool for forensic analysis carried out on proprietary software, and proprietary software Encase, used by law enforcement agencies, which, according to their website is “The Gold Standard in Forensic Investigations”. Encase is a software tool that is used for forensic investigations. It permits the investigator to acquire data from a wide variety of devices, complete a forensically sound investigation into the device, including finding things such as users of the device, dates of use, software and websites visited, and report on these findings. Another such proprietary programme is Access Data’s Forensic Toolkit (FTK), which can carry out very similar investigations. These digital forensics software solutions are designed for law enforcement and government agencies, and provide detailed user information, access to hidden files and content, email threads, file paths and image forensics. However, in the case of Autopsy, Encase and FTK, there is an expectation that the user will be trained in digital forensics and will be an expert in extracting relevant evidence, knowing what to look for and how to find it. In academic misconduct cases, it is highly unlikely that the investigating officer will be given access to the student’s (or contracted author’s) device, so whilst this information could be extremely helpful in determining the true authorship of a document, this is an unrealistic option for the average academic misconduct investigator. These tools and techniques are described in Johnson et al., (2022).

However, some options beyond text-matching tools and expensive and complicated software packages do exist to the investigator. Turnitin’s Authorship tool, for example, can already provide detailed reports on the device used to create and edit a document, including who the registered user of the device is, as well as the version of Word, number of revisions and so on. Whilst variances in these for a single document may have perfectly genuine explanations, they can certainly help indicate where further investigation and discussion with the student would be beneficial. This information can also be accessed without such expensive software, and forms part of the experimentation phase of this research.

Tools which automate identification of potential academic misconduct and provide evidence to support an investigation are of great value to assessors. As Prentice and Kinden (2018)
note in their analysis of paraphrasing tools, academics are often left in the position of taking on the role of detective, such that when they suspect a student of academic misconduct, they are forced to collect evidence, analyse motives and apply a ‘quasi-judicial if not criminological’ approach, and whilst academics may believe strongly that they only want to reward student work in a just way, they perhaps did not anticipate finding themselves in this role.

3.8.8 The role of detection

Research into the benefits of surveillance in reducing crime may be relevant here. Students who know there is a risk of being caught committing academic misconduct are less likely to engage in these behaviours. Lancaster (2022b) notes the importance of detection tools as a deterrent to students. This is reflective of research by Foucault (2020) who explored the role of surveillance in controlling imprisoned persons, primarily from a physical perspective, and of discussions from Galić et al., (2017) who noted that surveillance can be both controlling and also (possibly at the same time), protecting and caring. There is a deeper, philosophical consideration here too, around discipline. Foucault suggests that individuals internalise discipline and are less likely to break rules because they think they are being watched, even if they are not. Modern anti-crime specialists have taken this on board by creating posters of someone’s eyes in areas where theft occurs frequently (such as outdoor bicycle stands). The visual appearance of eyes makes thieves feel as though they are being watched and as such the crime rate is reduced and less bicycles are stolen (Nettle et al., 2012).

Turning to technology, a detailed review of literature relating to the impact of CCTV on crime is provided by Piza et al. (2019), who find that the use of CCTV was associated with significant reductions in both vehicle and property crime, with less effect on violent crime. Whilst these specific discussions are not directly relevant to academic integrity, the sentiment behind them is, demonstrating that the possibility of getting caught acts as a preventative measure when students are considering plagiarising or outsourcing academic work, whilst also potentially protecting students from essay mills touting for business through website blocking and tracking of Internet traffic. This adds to the body of knowledge that suggests that tools known by students to detect academic misconduct are a vital piece of the jigsaw.
3.9 Summary of Chapter 3

This chapter started by looking at existing research into student perceptions of academic integrity as a way of reviewing ways in which academic misconduct could be prevented or reduced (a good understanding of student perspectives helps academic institutions put appropriate measures in place to help reduce the temptation for students to act without integrity). The chapter went on to discuss digital forensics and where research into digital forensics exists, and to what extent this is related to academic integrity, as a way of looking at the next step of academic misconduct: that of detection. It examined what researchers currently know about the forensic artefacts that are automatically saved with a Word document, along with an introductory explanation of the OOXML format and artefacts that can be found when examining the decompressed files that make up a Word document. The relevance of these artefacts in relation to plagiarism and contract cheating was covered as an overview.

The existing tools for detecting plagiarism and contract cheating have also been discussed, as have the tools that are currently used in criminal investigations or investigations relating to intellectual property cases but it is clear that these subjects are not yet combined, so the tools that are currently used in the detection of contract cheating and plagiarism do not incorporate the forensic analysis of XML, and the tools that are used for investigating criminal or IP cases are not utilised for detecting contract cheating and plagiarism, and yet it could be said that plagiarism and contract cheating are both types of fraud.

All too often the focus of academic misconduct is on the final product. Existing research considers students’ views of academic misconduct, but usually through the lens of an assessor or experienced researcher and occasionally through self-selecting samples, rather than specifically through the eyes of the students themselves. Detection methods also fall short, focusing on the end product rather than trying to understand the process that has gone into creating that product, despite this being able to provide potentially rich information. Facilitating an automated method of detection, that goes beyond merely text-matching, and incorporates a variety of techniques could provide strong evidence for academic misconduct cases.

The next chapter will detail the findings from the primary research relating to students’ awareness of academic integrity and understand how important they feel this is before
moving onto the development of the detection tool in Chapter 5. This tool will repurpose
digital forensics methods and apply them to Word documents in order to reverse-engineer
them and discover how the document was created, providing a series of clues as to the
authenticity of a document.
Chapter 4: Results of the Primary Research into Student Perceptions

This chapter presents the findings from the two primary surveys on student perceptions. As described in the methodology (see section 2.5), two surveys were conducted at different points in the academic year. The students selected to take part were a convenience sample at the researcher’s institution and comprised one large group of engineering students, a smaller group of business students and a similarly smaller group of media students. The large engineering cohort accounts for the number of male respondents being greater than the number of female respondents, although the ratio of male to female respondents is much lower at survey point 2, where students were asked to complete the survey independently in their own time, rather than in class.

4.1 Results from the two surveys on student perceptions

4.1.1 Respondent numbers

![Respondent numbers chart]

Figure 7: Respondent numbers for the primary surveys

In total there were 90 respondents for phase 1 (out of a total of approximately 120 students in the classes selected on the day of the survey) and 34 respondents for phase 2 (out of a possible 191 students who were sent the link) with the lower response rate for the second survey being attributed to the method of delivery (online), rather than in person during a class as for the first survey. A high proportion of students (two thirds) live in the family home, which is considered to be due to the type of institution where the research took place, which
had a large population of local students at the time of the survey. It is noticeable that a larger proportion of female students completed the second survey. One of the cohorts was a large group of engineering students, with a smaller proportion of female staff. As the first survey was done in class, it is not surprising that the responses show a higher proportion of male students. As the second survey was done online, students were less likely to complete them and could be considered as self-selecting, and this may account for the larger proportion of female respondents.

4.1.2 Students’ perceptions of ‘plagiarism’

![Figure 8: Students' perceptions of plagiarism](image)

Following the question asking students ‘Do you know what plagiarism is?’ was a secondary question asking them to explain the term in their own words. Almost all students confirmed that they know what plagiarism is, and the follow up definitions demonstrated a good understanding of the term. Only one student at survey point 1 and no students at survey point 2 said they didn’t know what plagiarism is.
4.1.3 Students’ perceptions of ‘academic integrity’

Students were less certain of the term ‘academic integrity’ (Figure 9). In the first survey, just over 10% of the respondents said they knew what academic integrity is. Interestingly, those students who were not sure increased from the first survey point to the second, suggesting that students become more aware of the term, but less sure that they know what it means as their studies progress (bearing in mind these are first year students). This is reminiscent of Einstein’s famous quote – ‘The more I learn, the more I realise I don’t know’, i.e. that a little knowledge makes us better able to understand that we do not know everything. In this instance, it would appear that as students progress through their first year, they begin to realise that they do not fully understand academic integrity, which is the first step to gaining a better understanding. It also suggests that interventions very early in the year are likely to be less effective than towards the end of the year, when students may be more receptive to learning about academic integrity. Students who claimed to know what academic integrity is in this question were then asked to define academic integrity in their own words, with the following results:

- ‘Being honest’ (2)
- ‘Student Honesty’ (1)
- Attitude to learning (1)
- Being truthful in where information was sourced from
- Not cheating

Figure 9: Students’ perceptions of ‘academic integrity'
If you plagiarise then people in the academic field won’t think highly of you.

These definitions show that even the students who claim to know what academic integrity is do not have a full grasp of the meaning of the term.

4.1.4 Students’ perceptions of ‘contract cheating’

A surprising outcome from the surveys is that many students did not know what contract cheating was (Figure 10). Reading the definitions provided by the students who thought they knew the term contract cheating meant, it was clear that even those who said they understood the term tended to confuse it or use it synonymously with the term ‘plagiarism’:

- ‘When you don’t obey the rules of a contract you agreed upon’
- ‘Copying work’
- ‘Buying a service to write the essay’
- ‘Buying an assignment and using it yourself’
- ‘Plagiarism’
- ‘Stilling [sic] information from other students’
- ‘To take your class mate answers during exams’

Values in the next section are given for Survey Point 1 and Survey Point 2 in brackets (Survey Point 1 / Survey Point 2).
4.1.5 Students’ perceptions of the support they received

In general, most respondents felt that they were either very well supported by their teacher in their last piece of work (19 students - 21% / 9 students - 26%) or reasonably well supported (44 students - 49% / 18 students - 50%). Only 4 students - 4% / 3 students - 9% respondents felt poorly / not well supported.

4.1.6 Time keeping

When asked to consider the last piece of work the respondent had completed, none reported that they had handed in the work late or not at all at Survey Point 1 and only 1 respondent said they had handed in the work late or not at all at Survey Point 2. The majority of students stated that they had either plenty of time (44 students - 49% / 12 students - 35%) or handed the work in just in time (26 students - 39% / 20 students - 59%).

4.1.7 Cheating behaviours

When considering previous work, respondents were asked whether they had carried out any ‘cheating’ behaviours. Many respondents ticked several boxes in this category. Overall results showed that 24% / 41% had copied a sentence from the Internet or published sources without referencing, 13% / 16% had copied a paragraph or more from the Internet or published sources without referencing and 4% / 0% had asked someone else to do the work for them. These differences in self-reported ‘cheating’ behaviours varied the most from survey point 1 to 2 which may have been due to the reduced number of respondents and the likelihood that the more conscientious students would have completed the second survey point which was done outside of lesson time (i.e. in the students’ own time), though it is interesting to note the number of students who admitting to copying a sentence from the Internet or published sources at survey point 2. This survey question would have benefited from an option to state that they student had not carried out any of these behaviours: however, 43% / 35% of students did not tick any box, which could imply that they had not engaged in any of these behaviours. It should be noted that students could select more than one option in this question, so the percentages will not add up to 100%.

When asked if the respondent knew anyone who had cheated (by copying from other sources, or asking someone to do the work for them), 40 students - 44% / 25 students - 74% said they did not know anyone, 40 students - 30% / 25 students - 18% knew someone who had copied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

text, 6 students - 7% / 3 students - 9% knew someone who has asked someone else to do the work for them.

When asked in which situation it would be acceptable to get someone else to write an assignment for them, or to copy text from the Internet / printed materials, 73 students - 81% / 34 students - 100% said ‘Never’. For the first survey point, all other categories in this question (‘If you are stressed, if you run out of time, if you don’t think you are going to pass) received less than 6% of the responses.

4.1.8 Students’ awareness of academic misconduct penalties

![Survey Figure](image)

Figure 11: Students' awareness of academic misconduct penalties

Most students confirmed that they had the penalties for academic misconduct explained to them (49 students - 54% at survey point 1 and 23 students - 68% at survey point 2). Just under 20% of students at both survey points were not sure if they had the penalties for academic misconduct explained to them, which would suggest they could not remember what the penalties were, or that they considered they were not likely to engage in academic misconduct and therefore the penalties for doing so were irrelevant. Other questions on the survey showed that students were aware of where to go to for help if they were struggling with an assignment. However, this very much focuses on misconduct as opposed to integrity. It could also be argued that all students should be aware of the penalties for academic misconduct, not just the majority.
It is interesting that at survey point 1, 11% of respondents claimed to know what academic integrity was, but that this had reduced to 3% by survey point 2, which could potentially imply that students were more inclined to question their understanding later in the academic year.

At survey point 1, 57% of students said they had academic misconduct penalties explained to them, rising to 68% by survey point 2. Students at both survey points generally had a good understanding of what plagiarism was (96% and 94% respectively). These points reiterate the suggestion that most teaching staff (and indeed students) focus on the negative implications of academic integrity / misconduct, rather than the positive reasons for correct referencing and authentic work.

4.1.9 Students’ understanding of where to get help

![Figure 12: Students' understanding of where to get help](image)

Figure 12: Students' understanding of where to get help

50 students - 56% students at survey point 1 and 22 students - 65% at survey point 2 said they did know what to do if they were going to be unable to complete an assignment on time, 20 students - 22% / 6 students - 18% said they did not, and 14 students - 16% / 6 students - 18% said they were not sure.
Using Nvivo, it was possible to model various cross-referenced questions, which led to some interesting observations:

- The age group most confident in their study skills was the 18-19 year age group: 21 respondents in this age category felt ‘Mostly confident’ in their study skills, and 31 felt at least ‘neither confident nor lacking confidence, mostly confident, or really confident’. None of this age group lacked confidence.
- The category most likely to feel ‘really confident’ in their study skills was the 20-24 age group.

4.2 Summary of Chapter 4

Understanding student perspectives is vital for reducing plagiarism. Considerable research has been done into plagiarism and the many and varied issues surrounding this area, and suggestions have made as to how it can be detected and discouraged. Detailed qualitative evidence of student perceptions on their academic journey is more limited, particularly when reviewing where students start on this journey, and what helps or hinders them in their progress. This evidence is an important ingredient in developing a supportive, effective and successful learning experience. From the primary surveys carried out with students it was clear that whilst they had a good understanding of what academic misconduct and plagiarism is, they had a much weaker understanding of the term ‘academic integrity’, suggesting that the focus is primarily punitive, i.e., ‘do not do this or you will face consequences’, as opposed to positive, i.e., ‘it’s good to reference other people’s work when done correctly’. Also, students were less confident of their understanding of academic integrity at the touchpoint later in the year. Overall, they appeared to know what academic misconduct is and that they should avoid it, but that they did not have a good understanding of what academic integrity is and why the should ensure they use integrity in their work.

During the primary surveys, students were asked how they felt about students who cheated (either by copying from someone else or the Internet / printed materials, or by asking someone else to do the work for them), the majority of students who responded (35%) said that they didn’t mind that peers had cheated, although 12% claimed it made them angry and only 6% reported them. These students were also asked what might make someone breach academic integrity rules. Of the 18 respondents to this question, the most popular responses related to stress and / or pressure to do well along with struggling with the work and not
feeling able to complete it to the required standard. A smaller number (22%) said that students who cheat are lazy and want to complete the work with as little effort as possible. Again, this showed that students (and academics) tend to focus on what misconduct is, and not what integrity is. Institutional policies too often focus more on policies to prevent academic misconduct and the penalties that will be applied should a student be found to be engaging in these activities. At induction, students are warned of the dangers of outsourcing work, or copying or colluding, but they are rarely taught why proper referencing and citations are so valuable, and what can be gained from doing this correctly.

Given that students recognise that academic misconduct is a negative, the next phase of the research, forming the substantive part of this PhD, focuses on detecting academic misconduct when it occurs. Whilst deterrents such as creating authentic assignments and having clear penalties may help steer students away from academic misconduct, HEIs have a responsibility to ensure that students are awarded credit on the basis of their own work, and as such the focus moves to detection by exploring different ways of approaching detection of academic misconduct through a cyber security lens. Learning is a process of development, building on prior knowledge and using that to develop critical thinking skills. This process, when considered through the popular form of written assessment was discussed in section 3.4 in relation to academic integrity, and is explored in greater detail in 5.2 in relation to the evidence trail this process leaves behind. The next chapter details the way this evidence can be extracted and analysed to provide information on how a student has created their work, and whether this indicates that the work is authentic or not.
Chapter 5: Results of the Software Development of ‘Clarify’

This chapter details the investigations carried out in phase 2 of the research. During the technological investigations and development of ‘Clarify’, the five stages of Phase 2 discussed in the Methodology are followed. These include: Stage 1: Preliminary testing on a known case of contract cheating; Stage 2: Investigations into the forensic artefacts left in copy-and-paste plagiarism; Stage 3: More detailed investigations into the XML language; Stage 4: Development of a prototype tool with the working title ‘Clarify’; Stage 5: Developing the user interface and guided examples.

The chapter therefore contains the following sections:

- OOXML and the structure of a Word document, describing the file format itself, why it is designed in this way and what it comprises (section 5.1).
- The essay writing process and its relevance, including an introduction to RSID editing tags (section 5.2).
- Technological investigations which include the following:
  - An exploration of OOXML and RSID edit values in relation to a sample of contracted work (section 5.3, Stage 1).
  - An exploration of other XML forensic artefacts in copy-and-paste plagiarism (section 5.4, Stage 2).
  - A deeper analysis of forensic artefacts through a series of investigations (section 5.5, Stage 3).
- Development of ‘Clarify’, an academic misconduct detection tool using student submissions (section 5.7, Stages 4 and 5).

5.1 OOXML and the structure of a Word document

Microsoft Word uses the ‘Office Open XML Format’ (OOXML) format, which has been in use since Microsoft Office 2007 and was designed to bring several benefits to individuals, organisations and developers. These benefits include: improved damage recovery – because the various components of each document are stored separately, meaning that if one component is damaged it may still be possible to open the file; better privacy and control over personal information – because sensitive information can be more easily identified and
thus removed if required; and more compact – as the various instructions that make up the document are packed up into a single file. It is this feature that we can take advantage of when carrying out a digital analysis (Microsoft, 2019) as it is possible to unpack the file to explore its metadata. Some instructions that appear in a file are common across multiple documents (font formats for example), and others are specific only to the document in question (content and revisions).

A Word document is essentially a collection of other files compressed into a single ‘docx’ file, much like a zip file which contains a number of documents compressed for sending over the Internet. In very simple terms, this could be likened to a film, which contains a series of connected but separate elements, such as a script, musical score, orchestra, actors, special effects and more (Figure 13). Whilst the majority of people will only ever see the final product – that of the film itself – it is actually comprised of multiple elements that are linked together into a final product.

The XML files that make up the final Word document provide very useful information, such as metadata, document properties and time stamps, formatting, hyperlinks, uncropped images and the text itself, and yields very interesting data into how the document has been written. The feature that is particularly interesting in this metadata is the capture of edits carried out during the document writing process, using ‘Revision Save Identifiers’, or RSID values. This feature enables two or more authors to collaborate on the same document, with Word keeping an up-to-date version, tracking changes in real time. The RSID values are generated randomly, and according to ISO (2016) increment throughout a document’s life span, for

Figure 13: A film is a collection of other elements. The Last Jedi (The Walt Disney Company: image subject to copyright)
example when edits are carried out or a Save is actioned. However, testing in Stage 3 of the tool development established that the RSID values generated do not appear to increment throughout the editing process, or that if they do, it is not as simple as incrementing on the basis of edits carried out over a timeline. Nonetheless, the ability to extract and analyse these values can be very useful in establishing the structure of a document, whether it has ever shared ownership with another document or shows other signs of collusion, and whether it shows hallmarks of contract cheating or copy-and-paste plagiarism. By collating this metadata into a meaningful format, assessors can be given a summary of key information that could prove very useful in detecting and evidencing academic misconduct. For instance, a document creation date that precedes an assignment handout date would suggest that the student has used a previously written assignment from someone else as the basis of their response. An editing time that does not correlate with the length of the document would suggest that text has been copied from another source. Editing codes that match those of another student’s document suggest that the papers are from the same original source, indicating collusion and a lack of editing altogether may indicate contract cheating or plagiarism. In the contracted assignment discussed in section 5.3, preliminary investigations showed unusual editing patterns, which, when compared to an authentic student submission, showed strikingly different metadata.

It is simple to review the overt document properties of a file when opened in Word. This information can be found under the File menu by selecting Properties and then the Statistics tab, depending on the Word version (Figure 14). This information can be useful, but is not always reliable as it can be edited or deleted.
Instead, looking inside the packaged contents of this document, i.e., decompressing it, can reveal much more interesting data about the file and the way it was created. This can be achieved by changing the extension of the file from .docx to .zip, and then choosing Extract, or Unzip (depending on the operating system being used). Opening the folder that is created then reveals a series of subfolders, e.g., rels, docProps, word, and a single file called [Content_Types].xml as shown in Figure 15.

Within the extracted files for a Word document, there will be the following content as a minimum: _rels, docProps and word (folders); document.xml; settings.xml; styles.xml. It is possible that other files may exist within this structure depending on the content of the document (if tables and / or graphics are present for example). The [Content_Types].xml at the root of the folder contains a list of the content types of the parts within the package. The
The _rels folder tells Word how the parts relate to each other and to resources outside of the package. Rarely, the extracted files may not include the above, and this is discussed later during the investigations as being indicative of a non-native Word document.

The file containing most of the content is the document.xml file, and this file also contains the main RSID markup for the document, showing all the edit runs that still remain within the content and indicating how the document was built, as every edit that has taken place will sit inside an RSID tag. Within document.xml, styling is defined within the XML using a <w: namespace for Word documents, which has elements added to it in accordance with how the information is to be displayed. For example <w:document> tag tells Word that it is looking at a Word document, whilst <w:body> indicates that what follows is the main body of the document. Just like in HTML (Hyper Text Markup Language), used for building web pages, closing tags are used to mark the end of a run or block. Also similar to HTML, tags can be nested inside one another, so a <w:r> edit run tag may contain within it a <w:t> tag to indicate textual content.

Decompressing an example file called ‘test2.docx’ which contains just a few lines of text and some bold formatting results in the following files and folder when extracted (Figure 16):

![Figure 16: Decompressed docx file](image)
Opening the document.xml file into a web browser or text editor reveals the metadata of the file. In a very simple Word document (i.e., a simple file containing only one element or item, such as a single piece of text, a character or a line break) this file will contain the following structural elements (Table 3):

<table>
<thead>
<tr>
<th>document</th>
<th>&lt;w:document&gt;</th>
<th>The root element for the main Word processing document part</th>
</tr>
</thead>
<tbody>
<tr>
<td>body</td>
<td>&lt;w:body&gt;</td>
<td>The container for the main block level content</td>
</tr>
<tr>
<td>p</td>
<td>&lt;w:p&gt;</td>
<td>A paragraph</td>
</tr>
<tr>
<td>r</td>
<td>&lt;w:r&gt;</td>
<td>A run</td>
</tr>
<tr>
<td>t</td>
<td>&lt;w:t&gt;</td>
<td>Text</td>
</tr>
</tbody>
</table>

Each of these elements sits inside the parent element, but a parent can contain many child elements of the same name, so a <w:p> (paragraph) element must sit inside a body element, but a body element could contain several paragraphs.

A **paragraph** is one of the most basic units of block level content and defines content that typically begins on a new line. Each paragraph can contain three pieces of information: optional paragraph properties (that indicate how the paragraph should be formatted or styled and are identified by the <w:pPr> element), runs (or inline content, indicated by the <w:r> element, which demarcates a region of text) and further optional revision IDs (rsidR) which may be used to compare the contents of two documents (useful for shared documents, and version control for example).

After the paragraph element comes the **run** (<w:r>). The run demarcates an area of text with a common set of properties, where properties can be optionally controlled with an <w:rPr> run properties element.

Within both paragraphs and runs, unique revision identifiers (RSID values) can be found to track the editing session when the content was edited in some way (rsidR) or formatted (rsidRPr).

There are some existing resources which can be helpful in the forensic analysis of a document. Yakovenko (n.d.) is a software developer and blog post author, and provides a
clear outline of the docx structure. His online article assists developers in parsing the docx format for simple tasks such as indexing or converting to ‘txt’ format. Whilst providing a useful resource for understanding the format in more depth, there is no information about why the document is structured in this way, and what it can tell us about the construction of the document. In fact, Yakovenko states that “<w:rsidR> is an attribute that you can ignore; it’s used by MS Word internals”. The article is nonetheless helpful in terms of understanding the document structure.

Eric White (White, n.d.) is another blog writer, whose OOXML blog also includes some really interesting content. In particular, he references an OOXML tool extension for Chrome which allows for easy visualisation of docx files. It also has a comparison tool, where it is possible to add two files to the browser and have the OOXML tool compare them and highlight any differences. For a basic way to view docx files, this tool is excellent, and can be used to provide a very simple way for assessors to compare two students’ work for similarity. Exact matches of the sub folders and files are highlighted in green, whereas files that have changes are highlighted in red. The purpose of this tool is for developers working on OOXML formats where it is necessary to occasionally edit the file contents. The tool provides a quick way to unpack and analyse the file contents, and whilst not intended for use by academic institutions, it would be interesting to examine whether some of the techniques could be integrated with the author’s current research.

5.2 The essay writing process

Lancaster (2020) summarises that the writing process is made up of how someone writes, along with what they write (the content itself). In the how phase, where a student may be considering options, there will be bursts of typing, pauses for thinking and subsequent editing. Indeed, Trezise et al. (2019) suggest the use of keystroke logging via learning analytics to track this process as a way of detecting contract cheating, and this technique has been trialled by some exam proctoring services (Measure Learning, 2022). They suggest that keystroke logging allows for ‘analysis of the fluency and flow of writing, the length and frequency of pauses, and patterns of revision behaviour’. Throughout the process of putting an assignment together, there will be typical activities such as brainstorming ideas, researching, writing initial content, editing, deleting, adding citations, adding figures and tables, and finally proofreading and correcting. All this takes a considerable amount of ‘editing’ – i.e. activities inside the Word document itself. These edits are tracked within the
docx format using the RSID tags discussed above, and in an authentic assignment it is clear to see these RSIDs littered throughout the document as the author goes about the normal process of compiling their work. An example of the RSID editing tags is provided below, where a document is created containing the text ‘A simple document’ (Figure 17) and the XML that results from this is shown in Figure 18.

![A simple document](image)

Figure 17: Output of simple docx file with sentence added

This shows the visual output when viewed as a Word document. Below is the XML output of the document.xml file that is created.

```xml
<w:p w14:paraId="067BC715" w14:textId="7D9CA96B" w:rsidR="00606704" w:rsidRDefault="0072772B">
  <w:p>
    A simple document
  </w:p>
</w:p>
```

Figure 18: Basic XML markup showing rsidR editing tags

This shows the XML markup created by opening a blank document and adding a sentence of text. The RSID values in this case include w:rsidR="00606704" which marks the first edit run, and a default RSID value of w:rsidRDefault="0072772B" which is the default value applied to all editing runs within this block.

When an edit is carried out on the document, further RSID tags are added to show that a change has occurred. In the example above, the word ‘edited’ is added to the file and the XML extracted again to note the difference (Figure 19):
A simple edited document

Figure 19: Output of simple docx file with a word added

The changes to the XML following this edit are shown below (Figure 20):

Figure 20: Basic XML markup showing rsidR editing tags following addition of another word

In Figure 20 the addition of a further RSID value can be seen (w:rsidR="00A709BD") around the word ‘edited’ which indicates that it has been added at a separate time.

In a genuine work, the student will make many of these edits and alterations throughout the document writing process and as a result the document will contain many RSID values. Contrast this with content that is not created by the student themselves. If this content is from a contracted author, it is most likely that the student will copy the text into a new blank document before submitting. This results in stripping out almost all of the RSID editing, leaving behind a few useful clues to indicate that the work has been created in an unusual way. For example, a document.xml file containing large chunks of text with no additional RSID editing values tells us that this particular chunk of text must have been added to the document in one editing session, with no subsequent alterations – a challenging task for even the most competent author, even if they are a touch typist. If the student submits the contracted work in its original form, this too will provide clues for the assessor, such as a creator name and device that will not tally with the student who has submitted it. If the work
is copied and pasted from another source (for example, the Internet), again there will be large blocks of text without the RSID tags, but often these blocks of text require extensive reformatting, which again leaves useful forensic artefacts which can be suggestive of misconduct.

Further clues arise when the student goes through and carries out basic editing, such as adding their name and possibly the name of the institution they are studying at, or other individualised content. They may also amend work that they are not happy with, or to make it appear more authentic, perhaps by changing Americanised spellings to UK spellings throughout, or by changing reference to a location or subject throughout the entire document. These substitutions and edits are very clear to see when reviewing the XML of the document, as each edit run will be marked with an RSID value which changes at each save (whether autosave or a user invoked save). An authentic submission will, therefore, contain extensive RSID values throughout the document, with very few, if any, large runs of text (consider how difficult it is to write a large amount of text without making a single mistake and needing to later go back and edit it).

5.3 Stage 1: Using digital forensic techniques to detect contract cheating

In Johnson and Davies (2020a), a known case of contract cheating was examined and used as a case study for the application of digital forensics to detect academic misconduct. The following section details and expands on the work carried out and refers closely to the published article.

In this instance, there was no doubt that the work submitted by the student had been contracted out (or outsourced). The department that had issued the assignment was contacted by an essay writer who claimed that a student had requested some work be completed by them. This person (hereafter referred to as the Contracted Author) communicated on several occasions with the department, noting that the student who commissioned the work had a ‘habit of not paying after collecting the scripts’ (personal communication, 21 January 2018). When payment was not forthcoming, the Contracted Author searched for contact details of the university department issuing the assignment and made direct contact. They provided screen prints of the negotiations of the contract, along with a copy of the work they had provided but that had not been paid for. The assessors reviewed the work provided by the Contracted Author (hereafter referred to as the ‘contracted assignment’) and that of the
student, where it was clear that the works were very similar. As such, the process for dealing with Academic Misconduct at the institution was invoked, where the student in question confirmed that they had posted the assignment brief on an essay-writing website and subsequently asked one of the website authors to complete the work, which they subsequently submitted under their own name. Since the Contracting Author and the student both confirmed that the work was not the student’s own, the reliability of the information was not in question.

In the first instance, the XML of a researcher created file was explored to see what metadata it contained. The tests carried out in Johnson and Davies (2020a) are presented again here. A new blank document was opened in Microsoft Word 2016 v 16.0.4738.1000 and the text ‘Digital Forensics.’ was entered. The file was saved with the filename ‘Digital Forensics.docx’ and closed. The Word document was viewed in File Explorer and the file extension changed from ‘docx’ to ‘zip’, effectively telling File Explorer that the file is a compressed file, and as such it was possible to decompress (or ‘unzip’) the file using a zip tool. The attribute values in the document.xml file were reviewed and are described in Table 4, along with a snippet of code from the document.xml file (Figure 21).

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

Figure 21: XML code snippet from 'document.xml', Johnson & Davies 2020a

<table>
<thead>
<tr>
<th>Attribute Reference</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rsidRDefault</td>
<td>Default Revision Run Identifier</td>
<td>0009533F</td>
</tr>
<tr>
<td>rsidR</td>
<td>Revision Identifier for Run</td>
<td>00CE3196</td>
</tr>
</tbody>
</table>

Table 4: Attribute description, Johnson & Davies 2020a

Table 4 shows that a default rsidR value of 0009533F is applied to the document initially, and then as text is added, an edit run value or rsidR is created with the value 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 as shown in Table 5.

```xml
<w:p w:rsidRDefault="0009533F" w:rsidR="00CE3196">
  <w:r>
    <w:t xml:space="preserve">Digital</w:t>
  </w:r>
  <w:r>
    <w:t>Evidence</w:t>
  </w:r>
</w:p>
```

Figure 22: XML code snippet from 'document.xml' following an edit, Johnson & Davies 2020a

Table 5: Attribute description including additional value (Johnson & Davies, 2020a)

<table>
<thead>
<tr>
<th>Attribute Reference</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rsidRDefault</td>
<td>Default Revision Run Identifier</td>
<td>0009533F</td>
</tr>
<tr>
<td>rsidR</td>
<td>Revision Identifier for Run</td>
<td>00CE3196</td>
</tr>
<tr>
<td>rsidR</td>
<td>Revision Identifier for Run</td>
<td>00694EF2</td>
</tr>
</tbody>
</table>

The edit where the text change has occurred is identified by a new run value (rsidR) of 00694EF2. This attribute demonstrates that the rsidR attribute value alters as edits are made – though edits made in a single run (i.e. before a save or autosave) will share the same rsidR value, even if not directly next to each other. The original rsidR value is still visible because the word ‘Digital’ has not been altered during the second edit, and the rsidRDefault value still remains because there is at least one character within the document that matches from the first iteration of the document to the second. The greater the number of unique RSID values in a document, the greater the number of edits carried out. Where many appearances of a single RSID value occur, this indicates multiple edits were carried out across the document in the same editing session.
5.3.1 Digital forensic analysis

Some skeleton code was developed, written using PHP to provide an easy web interface for the results, and within this the test document ‘Digital Forensics.docx’ was analysed. The code was written to automatically extract the RSID values and display these alongside the textual content. Colour coding based on the RSID hexadecimal value was used to help visualise the RSID. Given that each RSID represents an editing session, in this example 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.

![Figure 23: First iteration of the tool to display RSID values (Johnson & Davies, 2020a)](image)

In Johnson and Davies (2020a), a paragraph of authentic work by the authors was also analysed using the skeleton code to explore what this should look like, given the usual editing process as described earlier. This is referred to as the Author Work Example Extract:

![Figure 24: Author Work Example Extract showing rsidR values (Johnson & Davies, 2020a)](image)

As would be expected, the text is littered with RSID values throughout as the authors write, revisit and change the work. Indeed, this paragraph alone includes twelve unique rsidR values for a section of text consisting of 106 words in total. Table 6 shows the ratio of edits to words.
Table 6: Author Work Example Extract ratio of edits (Johnson & Davies, 2020a)

<table>
<thead>
<tr>
<th>Words</th>
<th>Unique rsidR values</th>
<th>Ratio of edits to words</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>12</td>
<td>11.32%</td>
</tr>
</tbody>
</table>

Johnson and Davies (2020a) attempted to establish an order for these rsidR hexadecimal values as the ISO (2016) specification suggests that they are added to the document incrementally, but despite trying a variety of combinations including extracting pairs or octets of hex values and inverting them in various order, or reversing the values, it was not possible to confirm a timeline in this way. It would be interesting to explore this feature in future work as it may be that edit sequences are refreshed (resulting in a new sequence of RSID values being generated) when a document is opened and closed, or some other pattern emerges that would aid in developing a timeline of the content creation.

5.3.2 Analysis of copied work

Johnson and Davies (2020a) then experimented with sharing the document, as if one party had written the work on request, and then sent it to the other to use as their own. On receipt, the second party saved the document to their own file system, changed a single word (‘issues’ changed to ‘infringement’), and then resaved the work. Following this, the updated file was analysed by the skeleton software tool with the results shown in Figure 25.

Figure 25: Example of Author Work Example Extract following sharing with a second party (Johnson & Davies, 2020a)

In this extract there are now only 3 unique rsidR values despite the text being the same apart from just one word change. This gives the following ratio of edits to words, which is much lower than that of the original work as demonstrated in Table 7.
Table 7: Author Work Example Extract ratio of edits after sharing and editing (Johnson & Davies, 2020a)

<table>
<thead>
<tr>
<th>Words</th>
<th>Unique rsidR values</th>
<th>Ratio of edits to words</th>
</tr>
</thead>
<tbody>
<tr>
<td>106</td>
<td>3</td>
<td>2.83%</td>
</tr>
</tbody>
</table>

Colour visualisation of this provides more compelling results. In the example below, blocks of colour were substituted to represent text belonging to a single edit run or rsidR value. This yielded the following results (Figure 26 and Figure 27). Greyscale patterned versions of these images are available in Appendix B for clarity in print versions.

In the original Author Work Example Extract (Figure 26), there are many edits throughout the text, representing how the work was written and subsequently edited throughout. In Figure 27, the title is represented by a blue block of colour, and the body of the text sits within a purple block. The only other edit is a single red block, which is where the word ‘issues’ was changed to ‘infringement’. This shows how a large block of text was added to the document and required only one subsequent edit. The results suggest that visually representing work in this way would help an assessor very quickly identify where text has been copied and pasted, whether because it has been copied from an online source or taken from work provided by a contracted author.
5.3.3 Contracted assignment

Returning to the contracted assignment being analysed, the work written by the Contracted Author was compared to the work submitted by the student. This yielded the results shown in Table 8.

Table 8: Comparison of Contracted Author and Student Submission (Johnson & Davies, 2020a)

<table>
<thead>
<tr>
<th></th>
<th>Words</th>
<th>Unique rsidR values</th>
<th>Ratio of edits to words</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contracted Author</strong></td>
<td>3685</td>
<td>88</td>
<td>2.39%</td>
</tr>
<tr>
<td><strong>Student submission</strong></td>
<td>4172</td>
<td>15</td>
<td>0.36%</td>
</tr>
</tbody>
</table>

This data shows that the number of unique rsidR values reduces significantly in the student submission, and although the ratio of edits to words is lower than in the test case, it reduces from the Contracted Author’s work to the Student submission.

Visualising the contracted assignment in a similar way to the Author example in the previous section yielded similar results. In Figure 28, the Contracted Author’s work can be seen, littered with edits throughout (represented by different blocks of colour). In Figure 29, the student submission, there are very large chunks of colour, representing blocks of text added to the document in one editing session, with only small edits littered throughout. Both visualisations can be seen in pattern and greyscale in the appendix. Closer inspection of the XML of the student submission shows that these edits are to amend spelling errors, remove quotes around words, remove surplus spaces and similar. Statistical analysis of the data gives the results is demonstrated in Table 9.
Table 9 shows that the XML analysis of the Contracted Author’s work returns 88 unique rsidR values in the Contracted Author’s work – that is 88 separate editing sessions – compared to just 15 in the student submission. The ratio of unique edits to words helps to determine how many edit sessions have taken place in relation to the number of words within
the document – a high ratio is likely to be found in authentic work as the author writes and rewrites the document, revisiting the text many times. The total number of edits tells us how many edits were made overall within the document (remembering that a single edit value could appear several times throughout the document and comprise a number of alterations to the text, perhaps if the student changed all US spellings of a word to English spelling in one go, for example). Again, this shows that the contracted author carried out far more edits than the submitting student.

Whilst difficult to statistically evidence the unusual appearance of the work at this stage of the software tool development, the visual appearance highlights the difference between the originally edited work and the copied work. Aside from these samples, the authors also reviewed examples of their own original work and saw a similar pattern of editing as the Contracted Author’s, as opposed to what was seen in the Student Submission. Later developments of this technique reviewed better ways to calculate the statistics with more meaningful results, focusing instead on the average number of words per run and the average number of words per rsidR edit value. These yielded more useful findings that are discussed in detail in the next sections. Appendix E shows both the Contracted Author’s work and the Student Submission along with the ‘Clarify’ report.

5.3.4 Findings from the contracted assignment and next steps

This initial testing yielded promising results and suggested that further investigations and tests would be valuable. Areas considered worthy of further exploration included reporting on the frequency of minor edits, which could be indicative of spelling corrections, grammatical alterations and name changes, consistent with changes to contracted work, as well as testing the methods on copy-and-paste plagiarism. It might also be useful to highlight large blocks of unedited text to an assessor. Johnson and Davies (2020a) note that they had the luxury of having the Contracted Author’s work to compare with the Student Submission, which greatly aided the development of these forensic techniques, though this would not usually be the case in an academic misconduct investigation. They also noted the importance of alternative and genuine reasons for a document appearing with such a small number of edits because the student may have created sections of the document separately, and then copied and pasted them into a master document for example, or perhaps because they had carried out grammatical checks using an online tool that resulted in a final copy of their work that could then be presented for assessment.
5.4 Stage 2: Testing the method on copy-and-paste plagiarism

At this point, some very basic code (skeleton code) had been developed to automate the process of checking the document.xml file and displaying the results in a web browser. As discussed, this work has been developed around a known case of contracted cheating, which yielded certain forensic artefacts. This raised the question as to whether copy-and-paste plagiarism would also yield notable forensic artefacts, and if so, what they might be. A second stage of testing was therefore set up to review a series of documents, referred to as ‘sample files’, created on a range of different devices. The XML was reviewed to establish the forensic artefacts that could be extracted from work copied from external (or extrinsic) sources, reviewing not just the RSID editing tags but other artefacts that may be useful. This next section refers closely to the tests carried out on documents as researched by Johnson and Davies (2020b).

Three documents were created on Device A, one on Device B, four on Device C and two on Device D. The specifications for each device (operating system and software version) are detailed below (Table 10). These tests enabled further development of the code for the detection tool.

Table 10: Document and devices (Johnson & Davies, 2020b)

<table>
<thead>
<tr>
<th>Operating System</th>
<th>Word Version</th>
<th>Documents created</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device A: MacOS High Sierra V 10.13.6</td>
<td>Microsoft Word for Mac v 16.34 (20020900)</td>
<td>• Human Computer Interface Library</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Plagiarism Essay Wikipedia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Python Programming Essay Stack Overflow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Raspberry Pi Essay GitHub</td>
</tr>
<tr>
<td>Device B: Windows 10 Education v 1803</td>
<td>2019 MSO (16.10348.20020)</td>
<td>• Red Team Blue Team</td>
</tr>
<tr>
<td>Device C: Windows 10 Education v 1903</td>
<td>2016 (16.0.4954.1000)</td>
<td>• Securing the Internet of Things</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Computer Programming Wikipedia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Social Engineering Authentic</td>
</tr>
<tr>
<td>Device D: Windows 10 Education v 1903</td>
<td>Microsoft Word for Office 365 (16.0.11727.20222)</td>
<td>• Blockchain</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Online Discussion Fora</td>
</tr>
</tbody>
</table>
For each document, original text was added to the document which was then saved. Further text that had been copied from online sources was then added to the document and reformatted to match the style of the default document and remove hyperlinks and other extraneous information (footnotes, breaks and so on). The method of reformatting varied from document to document, in some cases by using the Format Painter (Paste Format) tool in Word, and in others by manually changing the font size and type. This reformatting replicates the actions of a student who has copied information from the Internet in order to make the work look consistent throughout. In some cases, this may be legitimate – if, for example, the copied text is a quote which is subsequently cited and referenced. The copied material used for the purposes of this testing are provided as a subsection of the references. A control document called ‘Social Engineering Authentic’ was also created using original text with a small amount of copied text that was subsequently paraphrased to replicate the actions of a student with a normal essay writing approach.

Once each document was saved, it was then converted to a zip file using the process previously discussed, and the XML document.xml file was manually reviewed in Notepad++ and Chrome to establish whether forensic artefacts relating to the process of copying and pasting could be found. Whilst the focus of the Contract Cheating case study had been on the edit values (RSID), this stage of the research focused primarily on other XML artefacts such as those relating to font, background, images and so on, to establish whether these could potentially be considered flags for plagiarised work.
The samples created by Johnson and Davies (2020b) are detailed below (Table 11):

Table 11: Documents created and device information (Johnson & Davies, 2020b)

<table>
<thead>
<tr>
<th>Document Name</th>
<th>Device</th>
<th>Source used</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Computer Interface Library</td>
<td>Device A (Mac)</td>
<td>Library – online journal articles</td>
<td>Two sources used</td>
</tr>
<tr>
<td>Python Programming Essay Stack Overflow</td>
<td>Device A (Mac)</td>
<td>Stack Overflow</td>
<td></td>
</tr>
<tr>
<td>Raspberry Pi Essay GitHub</td>
<td>Device A (Mac)</td>
<td>GitHub</td>
<td></td>
</tr>
<tr>
<td>Securing the Internet of Things</td>
<td>Device B (Windows)</td>
<td>Online journal article and blog post</td>
<td>Two sources used</td>
</tr>
<tr>
<td>Social Engineering Authentic</td>
<td>Device B (Windows)</td>
<td></td>
<td>Text copied for reference only, heavily edited or paraphrased</td>
</tr>
<tr>
<td>Red Team Blue Team</td>
<td>Device C (Windows)</td>
<td>Wikipedia and two online blog posts</td>
<td>Three sources used</td>
</tr>
<tr>
<td>Blockchain</td>
<td>Device D (Windows)</td>
<td>Web pages resulting from Google search</td>
<td>Two sources used</td>
</tr>
<tr>
<td>Online Discussion Fora</td>
<td>Device D (Windows)</td>
<td>Website and online journal article</td>
<td>Two sources used</td>
</tr>
</tbody>
</table>

As noted above, Johnson and Davies (2020b) created one genuine sample to act as a control (‘Social Engineering Authentic’). One of the most striking features of this sample was the
absence of the `<w:rFont>` element in the XML. In tests on other authentic works, the `<w:rFont>` element appears infrequently, to specify a single font attribute, e.g. `<w:rFonts w:eastAsia="Times New Roman">`. The exception to this is where the text is a field entry such as a Table of Contents or there has been a deliberate font change. It would not be seen on text that is formatted in the same way as the rest of the document. This is because when typing into a blank document, there would usually be no need to reformat the text to a different font, but when copying from another source, it is often necessary to reformat the text to match the default document.

In the other sample files created, the `<w:rFonts>` element appeared frequently, such as in the ‘Python Programming file’ for example:

```xml
<w:rFonts w:ascii="inherit" w:eastAsia="Times New Roman" w:hAnsi="inherit" w:cs="Consolas"/>
```

as well as in the ‘IoT’ file:

```xml
<w:rFonts w:asciiTheme="minorHAnsi" w:hAnsiTheme="minorHAnsi" w:cstheme="minorHAnsi"/>
```

Johnson and Davies (2020b) suggested that the more detailed and comprehensive `<w:rFonts>` tag could be because when text is copied from another source it contains a font specification from the source file as well as the new format style for the destination document. The Office Open XML glossary (ECMA, 2017) states that the attributes applied allow for the display of different subsets of Unicode characters that may not appear in the default character set including Asian or Arabic characters for example. These may therefore be a relic of language conversion, for example if a student had written the document in a language other than English and then needed to translate it for submission, and as such the appearance of these artefacts in the XML code may be entirely innocent.
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Figure 30: Frequency of <w:rFonts> element in sample files (Johnson & Davies, 2020b)

Figure 30 shows the frequency of <w:rFont> elements, which highlights the extensive font markup seen in the Python file, due to the programming code that appeared in the file. However, in some cases, the sample files containing copied text did not contain <w:rFont> elements. This appeared to be a result of using the Format Painter tool, which stripped out surplus font attributes. This was the case with the ‘Programming Wikipedia’ file.

Another element of interest during these tests was the appearance of the <w:rPr> element, which defines the properties for a specific run of edits, and may include attributes for font face as above (<w:rFont>), size (<w:sz>) and language pack (<w:lang>). One notable features of the <w:rPr> element is that it can suggest deletions or reformatting has occurred, rendering the formatting feature redundant. For example, <w:rPr> tags that contain various formatting features but no content (either text or graphics) are likely to be a relic of previous formatting. An example of a redundant <w:rPr> element is given below, showing that the font face is set along with a size and language, but that the tag itself is closed before any content is specified, thus rendering the tag redundant. In basic terms, this is like setting the font to size 16, bold and italics and then resetting it back to the default without actually typing anything. In the tests carried out, this feature was seen in several files including ‘Securing the Internet of Things’, ‘Raspberry Pi’ and ‘Python Programming’. Indeed, ‘Python Programming’ contained an extensive number of rPr elements compared with the other document, possibly because of the code snippets that it contains (Figure 31).
A few other features that were potentially of interested were explored. These included the `<w:sz>` element which sets the font size of the text and the `<w:shd>` element that sets a background style to the text. The ‘Python Programming’ file contained the most significant number of `<w:sz>` elements, likely because of the code snippets within the file, and the `<w:shd>` element was found in the sample files where reformatting had been done manually (not through the Format Painter tool), see Figure 32. Where original text is input into the document, there would be no need for this element to appear unless a shadow background was being applied, and therefore it was considered a flag for copied text where any background formatting needed to be removed by setting the background back to null or clear.
Another feature that Johnson and Davies (2020b) commented on is that in two of the sample files ‘Python’ and ‘Pi’ there is a much higher incidence of the elements and attributes previously discussed than in any of the other files. Indeed, an analysis of the XML words (elements and attributes) compared with the text word count (number of actual words) showed a much higher ratio (Figure 33). They noted that most of the samples demonstrate between 4 and 11 times the number of XML words to text words, whilst ‘Python’ and ‘Pi’ show 98 times and 219 times respectively. This is likely due to legacy formatting from the source file, along with reformatting required in the destination file. These features were further explored during the iterative development of the software tool to explore the value of these forensic artefacts in more detail.
Figure 33: Number of XML words in relation to document word count

5.5 Stage 3: Understanding the forensic artefacts

At this point a more detailed understanding was needed of how and why forensic artefacts appear in a document as it is created. It was clear that RSID edit values were created whenever an edit occurred within the document and that other forensic artefacts relating to formatting were potentially useful in establishing how the document was constructed. In order to establish how the edit (rsidR) and formatting (rPr) values are added to the document in an authentically written document, a series of tests were carried out as detailed below.
The following assumptions were made with a view to proving or disproving them during the investigations:

- An rsidR value is always generated when something is edited, even if the edit is just a formatting change.
- All paragraph elements include their own rsidR value.
- A `<w:r>` element always sits inside a `<w:p>` element.
- A `<w:t>` element always sits inside a `<w:r>` element.
- A `<w:t>` element always sits inside its own `<w:r>` element.
- The first `<w:r>` element in a `<w:p>` element has no rsidR value – it inherits the parent value.
- Text copied from elsewhere always has an rsidR value.

Within the XML format are the RSID codes which represent revision identifiers for style definitions. These take the format of a four-digit number that marks the editing session in which this style definition was last modified. Document content which shares the same RSID values indicates that changes were made in the same editing session.

**5.5.1 Investigations on a simple Word document**

The first series of investigations involved creating a new Word document and adding content, making simple edits and recording the changes to the resulting XML files. The steps taken for each investigation are detailed in Appendix C: Investigations, along with a table of observations, and key findings are documented here. Only the first and second investigations are shown in full in this section to outline the processes carried out.

**Investigation 1: Creating a simple Word document**

This test was carried out in Microsoft Word for Mac Version 16.40 (20081000), where the following process was followed:

- A new document was created in Word, using File, New Document.
- A title “Test 1” was added, and a new line created by pressing the Enter key.
- The document was saved with the file name “test1.docx”.
- The text “Content 1” was added, and a new line created (Enter).
• The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”, and a new line created (Enter).
• The document was saved again.
• The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed.

This yielded the following results:

Figure 34: Visual of how rsidR values are built in Investigation 1

The red bar | indicates a save point.
In table form, this is the process that was followed:

Table 12: Observations from investigation 1

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A new document was created in Word, using File, New Document.</td>
<td></td>
</tr>
<tr>
<td>• A title “Test 1” was added.</td>
<td>rsidRDefault: 00C90590 was generated</td>
</tr>
<tr>
<td>• The document was <strong>saved</strong> with the file name “test1.docx”.</td>
<td>rsidR: 00026756 was generated</td>
</tr>
<tr>
<td>• The text “Content 1” was added.</td>
<td>rsidRDefault: 00143CFB was generated and rsidR adopted the value of the previous rsidRDefault element</td>
</tr>
<tr>
<td>• The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”.</td>
<td>rsidRDefault remained as 00143CFB and rsidR adopted the value of the previous rsidRDefault element (which matches the current rsidRDefault value) No rsidRPr was generated as formatting carried out inline</td>
</tr>
<tr>
<td>• The document was <strong>saved</strong> again.</td>
<td></td>
</tr>
<tr>
<td>• The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was <strong>saved</strong> and closed.</td>
<td>New rsidRDefault: 00580F26 was generated (new save point) and rsidR adopted previous rsidRDefault value. rsidRPr took the current value of rsidR element. New rsidRPr value of 00580F26 (which came from the previous rsidRDefault value) added to the newly formatted words to show that the content had been edited / changed retrospectively – this suggests that the original text takes the rPr value to match the rsidR value for the run, and that the subsequent edit needs a new rPr value to show that it was edited separately, and therefore takes the rsidRDefault value for the run (so it’s different to the existing rPr value).</td>
</tr>
<tr>
<td>• <strong>FOOTER SECTION</strong></td>
<td>The header/footer section (w:sectPr) contained values for rsidR of 00143CFB (rsidR value from final &lt;w:p&gt; element) and rsidRPr of 00143CFB</td>
</tr>
</tbody>
</table>
The Footer section is added automatically to every Word document, regardless of whether it contains any information or not on the document itself.

For each investigation, any notable or key findings were extracted. For investigation 1, it was noted that when editing text that has already been written, the original text is given a \(<w:rPr>\) value to match the rsidR value for the run (\(<w:r>\)), and the new edit needs a new \(rPr\) value to show that it was edited separately, and therefore takes the rsidRDefault value for the run (so it’s different to the existing \(rPr\) value).

This first test inserted text prior to the first save point, so a decision was made to save the next experiment prior to adding any text to see what RSID values appear, if any.

**Investigation 2: Adding formatting to a document**

A second example was created called test2, but in this investigation, the file was saved with no content before any text was added. The following steps were carried out:

- A new document was created in Word, using File, New Document.
- The file was saved with the name “test2.docx”.
- A title “Test 2” was added, and a new line created (Enter).
- The document was saved again.
- The text “Content 1” was added, and a new line created (Enter).
- The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the word “inline”, and a new line created (Enter).
- The document was saved again.
- The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed.
This yielded the following results:

Figure 35: Visual of how rsidR values are built in investigation 2
To clarify this:

Table 13: Observations from investigation 2

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
</table>
| • A new document was created in Word, using File, New Document.  
• The document was saved with the file name “test2.docx”. | |
| • A title “Test 2” was added.  
• The document was saved again. | rsidRDefault: 00B63FEE was generated and rsidR: 00026756 was generated |
| • The text “Content 1” was added. | rsidRDefault changed to 00F7894 and rsidR adopted the value of the previous rsidRDefault element |
| • The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”. | rsidRDefault remained as 00F7894 and rsidR adopted the value of the previous rsidRDefault element (which matches the current rsidRDefault value) No rsidRPr required as formatting carried out inline |
| • The document was saved again. | |
| • The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed. | New rsidRDefault value of 00B635EF recorded (new save point) and rsidR adopted previous rsidRDefault value. rsidRPr took current value of rsidR element.  
New rsidRPr value of 00B635EF (which came from the previous rsidRDefault value) added to the newly formatted words to show that the content had been edited / changed retrospectively |
| • FOOTER SECTION | The header/footer section (w:sectPr) contained values for rsidR of 004F7894 (rsidR value from final <w:p> element) and rsidRPr of 004F7894 |

This showed that carrying out a save before entering any content does have any significant impact on the rsidR and rsidRDefault information.
Other notable findings were:

- The opening rsidR value is the same for both documents (first `<w:p>` element).
- The rsidRDefault value in each paragraph becomes the rsidR value for the next paragraph.
- Where content is reformatted after typing, an rsidRPr value is inserted into the parent `<w:p>` element, inheriting its value from the preceding `<w:p>` element (matching the rsidR value in the current `<w:p>` element).
- The run that contains reformatted content generates a new rsidRPr element that inherits its value from the parent rsidRDefault element.
- The footer contains an rsidR value which is inherited from the final `<w:p>` element of the document.
- The footer contains an rsidRPr value which is inherited from the rsidR value of the final `<w:p>` element in the document.

**Investigation 3: Editing a document**

This test involved going back to an earlier edit point in the document (which had already been saved) and making a further edit to review the impact this has on the RSID values. A new document was created in Word, using File, New Document

All observations made following Test 2 still held true. Namely:

- The opening rsidR value is the same for all three documents (first `<w:p>` element).
- The rsidRDefault value in each paragraph becomes the rsidR value for the next paragraph.
- Where content is reformatted after typing, an rsidRPr value is inserted into the parent `<w:p>` element, inheriting its value from the preceding `<w:p>` element (matching the rsidR value in the current `<w:p>` element).
- The run that contains reformatted content generates a new rsidRPr element that inherits its value from the parent rsidRDefault element.
- The footer contains an rsidR value which is inherited from the final `<w:p>` element of the document.
The footer contains an rsidRPr value which is inherited from the rsidR value of the final <w:p> element in the document.

Several variations on these tests were carried out. In addition to the tests documented in the Appendix, the following tests were carried out:

- Immediately following the addition of the final sentence and before a save was invoked, an edit was carried out to an earlier part of the document.
- Immediately following the addition of the final sentence and before a save was invoked, a word was deleted from earlier in the document.
- Immediately following the addition of the final sentence and before a save was invoked, several one-word edits were carried out on non-consecutive parts of the document.

Through this testing, it was possible to confirm the following:

- An edit carried out to the document before a save adopts the rsidRDefault value of the <w:p> element that precedes it in time (not related to location within the document).
- When formatting is altered retrospectively, the text retaining the original formatting is allocated an rPr value that matches the rsidR value of the run (<w:r>), and the text with the new formatting is given a new value for rPr based on the current rsidRDefault value of the run.
- When an edit occurs after a save, it is given a new (unique) rsidR value.
- Edits carried out in the same sessions not only bear the same rsidR value, but they also take the value from the preceding (in time) edit’s rsidRDefault value. Deletions do not show a new edit value.

These were important observations. They showed that it may be possible to identify edits carried out in a single session (i.e. between save points), regardless of whether they were consecutive or not within the document. Finding that the value of rsidR given to an edit is also inherited from the previous (in time) run’s rsidRDefault value could also be used to help determine the editing sequence of a document, although this will quickly become very complex, as was found in the following investigations.
Investigation 4: Copy, Paste and Edit

In the next investigation, the content from test3 was copied and pasted into a new file called test4. Some minor edits were carried out in the new document. This resulted in the addition of an rsidP value in each paragraph (<w:p>). This was not present in the original version, but was added to indicate that the paragraph was edited (added to the document) and the value allocated to this element 009B0641 indicates that all paragraphs were added at the same point in time (during a single edit, between any save points). This new rsidP value did not appear in the final (new) <w:p> element.

There were new tags of <textId="7777777"> inside all but the first <w:p> element. This appears to be as a result of copying and pasting the text from another document.

All paragraphs had the same rsidRDefault and rsidR values, which suggests that when information is copied from one file to another, most of the original rsidRDefault and rsidR values are removed and replaced with a single value throughout. This would correlate with the knowledge that the text was added in one go to the document.

One line (where the text had been reformatted in the original) retained some original information from the first file for its rsidRPr formatting elements, showing that legacy formatting changes may be retained.

There was a final (empty) paragraph which had the rsidR value of 00026756 – this was the original opening rsidR value from the original document.

5.5.2 Investigations to explore the editing sequence

The next few investigations explored the ordering of rsidR values within each document.

Investigation 5: Exploring the sequence of document edits

In the preceding examples, it seemed that edits adopted the rsidRDefault value of the paragraph which was part of the previous edit. The next investigations aimed to test this theory. In investigation 5 there was definitely a sequence for the editing. As suggested from earlier investigations, the rsidRDefault of a paragraph appears to become the rsidR of the next edit to be carried out in many cases, but at this point it could be due to the artificial way the document was created. It is also possible to see in this example that two edits were carried
out in separate places within the document, but between save points (the addition of the title ‘Plagiarism Experiments’ and the text ‘This phrase was inserted retrospectively’ on the Content 3 line. Both these additions bear the same rsidRDefault value to show that they were carried out in the same editing session.

**Investigation 6: Saving an existing file with a new file name**

The file test5.docx was reopened and the text “Student Name” added underneath the title ‘Plagiarism Experiments’. It was then resaved as test6. The only effect this had on the XML was to introduce a new (second) paragraph element for the text ‘Student Name’ with an rsidRDefault value of 000C2952 and an rsidR value of 000C2952 (both values were unique to the document and not inherited from anywhere else).

**Investigation 7: Multiple edits in a single editing session**

A new editing session was introduced into test6.docx. The purpose of this was to review the effect on the XML of multiple edits before a save. test6.docx was reopened and saved as test7.docx. Each appearance of the word ‘Content’ was replaced with the word ‘Heading’ in one editing session. The text ‘Test 5’ was edited to ‘Test 7’. At the end of the document, a final sentence was added with the text ‘End of document’. The document was saved.

From this test, it was possible to conclude that edits carried out in the same session (i.e. between saves) will bear the same RSID values, whether that is in the rsidRDefault value or the rsidR value. Again, it was found that content added sequentially in the document results in the rsidR value for the run inheriting the rsidRDefault value from the preceding edit.

**5.5.3 File actions and further editing investigations**

**Investigation 8: Effect of file actions on the XML**

A new document was created and saved with the filename “test8a.docx”. Using the File Manager, this was duplicated and saved with the filename “text8b.docx” and the document.xml file was checked to ensure that this has no impact on the resulting values (which it does not). test8b was reopened and random text was added to the document using =rand() to generate text automatically, and the document saved. Again, the file was duplicated using File Manager and the duplicate copy given a name of “test8c.docx”. test8c
was reopened and five minor edits carried out and the XML reviewed for all three documents to facilitate drawing conclusions around the rsidR and rsidRDefault values. This was repeated in Investigation 9 with the same results. The findings were that copying (or duplicating) a file using File Explorer has no impact on the document.xml file (i.e. no rsidR / rsidRDefault values are changed and the two document.xml files are identical.

The document test8b showed that the content added in one edit using =rand() resulting in each paragraph having matching rsidRDefault and rsidR values. The only paragraph that was different was the final paragraph, which inherited its rsidR value from the previous saved version of the document (test8a.docx). text8c also confirmed previous findings that all edits (except deletions) carried out in the same editing session bear the same (new) rsidR value.

5.5.4 Summary of findings from Investigations 8 and 9

Key findings from these investigations include:

- A document created and saved multiple times (with no changes other than to the file name, carried out in File manager) will retain the same rsidR and rsidRDefault values throughout.
- Adding random text in one edit results in a new rsidR and rsidRDefault value throughout, except for the last paragraph containing text, which takes its rsidR value from the previous document.

5.5.5 Entering text in a single session

The next two investigations include manually entering text into a document and not carrying out any user-invoked saves to observe the effect on the XML.

Investigation 10 – 13: Further exploration of rsidR values

Text was typed manually (3 paragraphs) into a blank document with no saves during the addition of the text to review the resulting RSID values. This was repeated in investigation 11 with the same results.

A new finding emerged at this point. Having completed a number of investigations it became apparent that the initial creation of a file appeared to include a device identifier: this is an rsidR value that was specific to the device it was created on. Later testing showed that this
device identifying value remains with the first content added to the document, although without knowing the device identifier from the outset it would be difficult to extract the value specifically, because the value is in the same format as all other RSID values. Several researchers have confirmed that documents shared across multiple devices will share a common rsidR value (indicating potential collusion in the case of Academic Misconduct) and the availability of a device identifier may also be useful.

These tests also confirmed that edits carried out in the same session (between saves) will have the same rsidR values.

**Investigation 14 – Mapping the rsidR values (TestF)**

Tests were carried out (TestF1, TestF2, TestF3 and TestF4) to map the evolution of rsidR values in a longer document. A combination of manually typed additions to the file, as well as insertions of random text using =rand() and copied and pasted text were used. A full table of the observations and RSID sequencing is provided in the Appendix.

By the end of the test, when the XML was reviewed, it was found that two clear line spaces had been added at the end of the document. These were not deliberately added. However, reviewing their rsidR values explains where they originated from. The first clear line space bears an rsidR value of 005F4050, which belongs to edit session 4. This then appears to be an accidental insertion of space after the final text was pasted in. The final clear line space bears an rsidR value of 002C66F6 (which is also the formatting rsidRPr value). This matches the first edit session rsidRDefault value, meaning that the clear line space (CLS) at the end of the document was actually inserted when the document was first created and has just been ‘pushed’ down the page towards the end of the document. This highlights that legacy RSID values are likely to occur when files and text are shared.

**5.5.6 Preliminary summary of findings from these investigations**

In terms of the assumptions made at the start of these investigations, the following were confirmed or disproved as indicated in Table 14:
Table 14: Summary of findings from the investigations

<table>
<thead>
<tr>
<th>Description</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>An rsidR value is <strong>always</strong> generated when something is edited, even if the edit is just a formatting change.</td>
<td>Investigations show this to be the case. Although it should be noted that the rsidR value may not be a new value, depending on the editing session in which the edit took place.</td>
</tr>
<tr>
<td>All paragraph elements <code>&lt;w:p&gt;</code> include their own rsidR value.</td>
<td>Technically it is possible for a <code>&lt;w:p&gt;</code> tag to exist without an rsidR value in terms of the XML coding structure. However, the likelihood of this actually occurring is very slim (and was not found in any of the investigations). This is because <code>&lt;w:p&gt;</code> is the paragraph level or main block-level container in the document, and simply by virtue of opening a document and placing the cursor in it means an edit of some sort has occurred and will therefore introduce an edit value (rsidR) to the document.</td>
</tr>
<tr>
<td>A <code>&lt;w:r&gt;</code> element always sits inside a <code>&lt;w:p&gt;</code> element.</td>
<td>Since <code>&lt;w:p&gt;</code> is the main block-level container, the <code>&lt;w:r&gt;</code> element, which is a non-block-level (or in-line) container will always sit inside the <code>&lt;w:p&gt;</code> element.</td>
</tr>
<tr>
<td>A <code>&lt;w:t&gt;</code> element always sits inside a <code>&lt;w:r&gt;</code> element.</td>
<td>This appears to be true.</td>
</tr>
<tr>
<td>A <code>&lt;w:t&gt;</code> element always sits <em>inside its own</em> <code>&lt;w:r&gt;</code> element.</td>
<td>This appears to be true, and is presumably because the creation of <code>&lt;w:t&gt;</code> occurs when text is added or edited and would only be split into multiple text runs where there are formatting changes or later additions / edits / deletions, all of which would incur an updated rsidR wrapper. However, it is possible for the rsidR value itself to appear multiple times throughout a document (i.e. there may be multiple runs or <code>&lt;w:r&gt;</code> elements with the same rsidR value in a single document).</td>
</tr>
<tr>
<td>The first <code>&lt;w:r&gt;</code> element in a <code>&lt;w:p&gt;</code> element has no rsidR value – it inherits the parent value.</td>
<td>Tests indicate this to be true, though it is likely that the rsidR value for the <code>&lt;w:p&gt;</code> element will update when an edit is made to the run inside it, and subsequent edits inside the same <code>&lt;w:p&gt;</code> element will then generate new runs / run values.</td>
</tr>
<tr>
<td>Text copied from elsewhere will always have an rsidR value.</td>
<td>Yes, though typically blocks of text copied from elsewhere (whether moved within the existing document or from another document or resource) will share the same rsidR value, even if there is other XML markup surrounding the text.</td>
</tr>
</tbody>
</table>
Other interesting information established through these investigations includes that the first rsidRDefault value in a document relates directly to the device and software on which the document was first created. In early investigations, the software used was Microsoft Word for Mac Version 16.40 (20081000) running on MacOS High Sierra V 10.13.6. All these investigations have an initial rsidR value of 00026756 and this value is seen throughout Investigations 1, 2 and 3. At this point the device was replaced and investigations carried out carry an initial rsidR value of 00AA7542. Later testing, when the investigations were repeated on MacOS Big Sur V 11.7, carried an initial rsidR (device identifier) value of 00606704.

In relation to edit sequencing within a document, it was found that an edit will take its rsidR value from the preceding edit’s rsidRDefault value. It was also found that edits carried out in a single edit run will share the same rsidR values, even if they are not consecutive in location.

It proved highly challenging to try and establish a sequence of editing using rsidR values, despite the ISO standard suggesting these values are added incrementally. Several sources on XML also note that ‘A producer may choose to increment the revision save ID to indicate subsequent editing sessions to indicate the order of the modifications’ (ECMA, 2017) but no other details could be found on how to achieve this, or indeed who the ‘producer’ refers to, though it may refer to other uses of XML (programming for example) and not to Word documents per se. Focus was therefore moved to other interesting and useful forensic artefacts.

All investigations were repeated multiple times to ensure that the instructions were repeatable with the same findings. As the investigations progressed, skeleton code of the detection tool was being developed. This enabled investigations to be reviewed both manually and within the tool.

### 5.6 Further exploration of the unzipped folder

As noted previously, a Word document is comprised of a number of other files, which when compressed result in a single Word document file for submission by the student. Most students will be unaware that a document is actually a group of documents combined together and would never have cause to examine the underlying files, as all of this is done within the Word processing software itself. Thus far, only the document.xml file has been examined, but other files within this structure also contain interesting and useful artefacts that may be useful
in detecting academic misconduct. The files within the decompressed folder are detailed in the following subsections.

5.6.1 document.xml

This file contains the main body of the document, including all the text. This file has been explored in the previous sections.

5.6.2 core.xml

This file contains the core properties for the document, such as the document title, who created the document, who last modified the document and the number of revisions of the document. Anomalies in author names, dates and revision number could be indicative of academic misconduct.

5.6.3 settings.xml

This file contains a list of all the rsidR values that have ever appeared in the file. This includes rsidR values that have appeared in the document at any point, even if all the contents have been deleted or the document completely blanked (Didriksen, 2014; Fu et al., 2011). A substantial difference in the number of rsidR values in settings.xml compared to document.xml (which only retains current RSID values) could indicate significant deletions have occurred. It is also of interest for checking against other student submissions (indicating possible collusion), as any matches of RSID values would be highly unlikely to occur by accident, and would instead indicate that the two documents originated from a common document at some point, even if the contents differ.

5.6.4 app.xml

This file contains the Application-defined properties for the document, such as Word version, the total time spent editing the document and the word count. In the investigations phase of the research though, it became apparent that the word count given in app.xml is not an accurate reflection of the total number of text words in the document. To test this, three simple files were created called Word_count_check.docx, Word_count_check16.docx and Word_count_check17.docx. The results of these investigations are given in Table 15:
Table 15: Word count results from app.xml

<table>
<thead>
<tr>
<th>Document name</th>
<th>Word count from app.xml</th>
<th>Actual word count given in Word GUI and manually checked</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word_count_check.docx</td>
<td>11</td>
<td>15</td>
</tr>
<tr>
<td>Word_count_check16.docx</td>
<td>12</td>
<td>16</td>
</tr>
<tr>
<td>Word_count_check17.docx</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>

To further check this, some authentic files were also analysed, which yielded the following results (Table 16):

Table 16: Word count results from app.xml on longer samples

<table>
<thead>
<tr>
<th>Document name</th>
<th>Word count from app.xml</th>
<th>Actual word count given in Word GUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>L3001</td>
<td>1387</td>
<td>1311</td>
</tr>
<tr>
<td>L7001</td>
<td>1138</td>
<td>1190</td>
</tr>
</tbody>
</table>

There appear to be a number of possible reasons for this anomaly. It could be that whilst Word will display an accurate word count in the GUI, unless the document is forcibly updated prior to exiting, this value does not appear to update. To test this theory, two documents were created with the same text consisting of five words in total. In the Word GUI, the word count was accurately given in the bottom left corner. In the first document, the file was simply saved and closed after the text was added. In the second document, the word count was manually checked by selecting the Tools menu in the GUI and selecting Word Count. This accurately showed that the document had 5 words. However, reviewing app.xml for both files and later in the detection tool ‘Clarify’ described in Chapter 5 yielded the results shown in Table 17:
This is further complicated by the inclusion of numbers and line breaks, all of which can complicate the counting. For this reason, a coded word count was added as part of the programme by adding the contents of all <w:t> tags together. Whilst this is not guaranteed to be completely accurate (as it does not accurately deal with line breaks and numbers) it is closer to the actual number of words in the document and is deemed sufficiently close to be useful in the investigations and subsequent ratio calculations. However, this would be worth further scrutiny in later iterations of the tool.

### 5.7 Stage 4: Testing ‘Clarify’ on student submissions

The various aspects of document unzipping have been explored in some detail, and initial findings from the skeleton code of ‘Clarify’ provided interesting and valuable analysis of some documents. However, tests by this point had only been conducted on fabricated work (i.e. samples created by the author, plus one case of contract cheating), and testing on ‘real’ student work was the next appropriate step for software testing. To further examine the value of this approach, a sample of 10 student submissions known to be in breach of the academic integrity policy at a Higher Education institution (i.e. having been through an academic misconduct board and been found in breach) were analysed using the tool, to help establish a baseline of artefacts. This baseline could be used in future developments to highlight potential flags requiring further investigation.

Submissions belong to a variety of course levels, from Level 4 (first year undergraduate) through to Level 7 (postgraduate). Submissions are encoded in the order they were analysed, with Ln to represent the level (where ‘n’ represents the level – e.g. L4 is a Level 4 submission). The remaining numbers are simply sequenced incrementally.
The following student submissions were used during this iteration of ‘Clarify’, all of which were submitted by students who had been found guilty of plagiarism offences. These submissions focus on plagiarised work with high Turnitin scores, partly due to the difficulty in getting samples of contracted work:

Table 18: Student submissions for testing

<table>
<thead>
<tr>
<th>Submission Reference</th>
<th>Turnitin similarity score</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4002</td>
<td>49%</td>
<td></td>
</tr>
<tr>
<td>L4003</td>
<td>45%</td>
<td></td>
</tr>
<tr>
<td>L4004</td>
<td>42%</td>
<td></td>
</tr>
<tr>
<td>L4005</td>
<td>79%</td>
<td>In class test</td>
</tr>
<tr>
<td>L4008</td>
<td>53%</td>
<td></td>
</tr>
<tr>
<td>L5001</td>
<td>50%</td>
<td></td>
</tr>
<tr>
<td>L5002</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>L7001</td>
<td>51%</td>
<td></td>
</tr>
<tr>
<td>L7003</td>
<td>46%</td>
<td></td>
</tr>
<tr>
<td>L7004</td>
<td>74%</td>
<td></td>
</tr>
</tbody>
</table>

The samples tested were coded according to their level, with L4 representing a first year undergraduate assessment. The number (e.g. 002) refers to the sequence of testing only. Not all samples were included as some failed to open in ‘Clarify’, particularly during the early stages of testing when the tool was still in development, due to anomalies in the coding which were later rectified. Other files were not included because they had been opened by a member of staff (as recorded in “last modified by”) and as such were considered potentially corrupt.

Various artefacts were recorded during the analysis: these included the RSID values, along with other XML features that had been noted at various stages of the research to this point, and that were deemed to potentially be useful in determining the authenticity of the document.
The following artefacts were recorded for each sample (Table 19):

Table 19: Artefacts recorded for sample documents

<table>
<thead>
<tr>
<th>Word count</th>
<th>How many words there are in the submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revisions</td>
<td>Number of revisions carried out on the document by one or more authors (includes things like insertions, deletions, changes to runs and markup).</td>
</tr>
<tr>
<td>TotalTime</td>
<td>The length of time the document was open for editing</td>
</tr>
<tr>
<td>rsidR</td>
<td>Number and value of revision identifiers</td>
</tr>
<tr>
<td>w:rFont</td>
<td>Font defined for a specific run</td>
</tr>
<tr>
<td>rFont: w:ascii</td>
<td>Font definitions for ascii values in a specific run</td>
</tr>
<tr>
<td>w:14 textID with a value of 77777777</td>
<td>The w:14 textID attribute sets a value for text runs (although is not always present). A value of 77777777 appears to be a default value, and was noticed frequently alongside copied and pasted text in early testing</td>
</tr>
<tr>
<td>w:shd with a value of clear</td>
<td>Appears where the background has been changed to clear</td>
</tr>
<tr>
<td>w:sz</td>
<td>Appears where the font size has been adjusted</td>
</tr>
</tbody>
</table>

An explanation of the elements and any assumptions included in this study are detailed in the following subsections.

5.7.1 Revisions

The value ‘Revisions’ found in the core.xml file relates to the number of times the document has been saved or ‘revised’. Limited testing on new documents confirmed this, but if a document was re-saved with a new name, the number of revisions reduced, indicating that this value may not be fully reliable.

5.7.2 TotalTime

The TotalTime element ‘Represents the number of minutes that the document has been open for editing since it was created.’ (Microsoft, 2010). This should not be confused with time spent actually editing a document, as a document can remain open for a long time with no actions being carried out on it. However, a very short TotalTime does indicate that very little time was spent editing.
5.7.3 rsidR

Following the previous investigations, the rsidR element is a useful value in analysing how the document has been edited. By way of a reminder, edits that are carried out in one editing session will bear the same rsidR value. Hence a document that carries very few unique rsidR values suggests that only a very small number of editing sessions have taken place – this would be highly unusual for an authentic student submission. rsidR values are gathered from both the document.xml file (which contains the text of the content) as well as from the settings.xml file. The settings.xml file contains all rsidR values that have ever existed in the document, even if the edited content that caused the element to be added has subsequently been deleted from the document. In addition, ‘Clarify’ records the number of unique rsidR elements. According to Joun et al. (2021) RSID values are created with the same rules on all versions of Word from 2007 onwards, and the operating system bears no relevance to this process. The author experimented on multiple different operating systems as can be seen during the various investigations and concurred with this view. Therefore, all tests carried out should be replicable across different versions of Word running on different operating systems.

5.7.4 rFont

rFont specifies the fonts which will be used to display the text contents of a run. If not present, the parent value (or default value for the document) will be used. The property w:ascii is to determine the font for all characters in the ASCII range 0-127 (the range of printable characters in the letters “a” to “z” and numbers “0” to “9”). The property w:cs relates to characters that are within the Complex Script range. By way of example the code

<w:rFonts w:ascii="Arial" w:cs="Times New Roman"> will display any characters in the ASCII range in Arial font, and any characters in the Complex Script range as Times New Roman, for this run only, overriding any default font settings. The assumption here is that unless the font actually differs from the default (perhaps for a heading, or table of contents entry for example), there should be no need to specify its setting. Hence font changes that simply result in the text being displayed in the default font could indicate that the text has come from elsewhere and was originally in a different font.
5.7.5 W14 textID

The w:id attribute sets a value for text runs (although is not always present). A value of 77777777 appears to be a default value and was noticed frequently alongside copy-and-paste text in early testing.

5.7.6 w:shd

This element specifies the background colour of content. Where text is written in the default font this element will not appear. If text appears on a different coloured background (highlighted, for example, or in a block heading) then w:shd will appear in the XML. The presence of w:shd where the text appears to be in its default font suggests there was a need to revert back to the default appearance. This typically happens with text copied from the Internet, though there may be other reasons why this formatting has been necessary.

5.7.7 w:sz

This element appears where the font size has been changed. The assumption here is that where the font is in the default size, ‘w:sz’ has been required to return it to this setting as it was in a different size somewhere else (and flagging potential copy-and-paste violations).

5.7.8 rsidRoot

Whilst not initially tested in the earlier investigations, this element is interesting to explore. It is the first RSID value to be allocated to a document. It marks the first edit carried out and is stored in the settings.xml file. Where documents are copied, they will bear the same rsidRoot value, even if all contents are deleted. Therefore, two documents bearing the same rsidRoot value must have originated from the same common document. This can be useful in collusion cases.

5.8 Tool development

The student submissions detailed above were tested in ‘Clarify’. ‘Clarify’ is a web-based tool, written in PHP. Firstly, the file being examined is processed into a PHP Dom Document, extracting the contents of the compressed file for analysis. The Document Object Model (DOM) is an object-oriented representation of a web page, facilitating extraction and analysis of contents. The file being examined is passed to the programme, which converts it
to a DOM document, and PHP is then used to extract more detailed artefacts from the submission. The results of the analysis are then displayed in a web browser.

5.8.1 Analysis of the document.xml file

During this iteration of code development, the formatting of the visual output was improved. The text itself can be displayed on the output page, or it can be displayed as colour-coded blocks, with each colour representing a single editing session. For example, if the word ‘help’ was changed to ‘assist’ throughout a document, with the change being made in a single editing session (e.g. via Find and Replace), the colour blocks representing the word ‘assist’ would bear the same colour. In addition, text runs may be broken up by grammatical errors, which are marked by Word for checking.

![Colour-coded visualisation of text showing a normal editing pattern](image)

Figure 36 shows the text runs of one of the student samples separated into blocks, with text belonging to a single edit run bearing the same colour. A greyscale version is available in the Appendix, which is easier to read in print versions of this thesis. This visualisation indicates a fairly standard pattern of editing, where there are some text runs together within the same edit run value, and smaller retrospective edits embedded within (as the author updates sections and makes adjustments and amendments). Authentic work would usually bear many colours when displayed in this way because of the way a typical document is written, as discussed in the author’s previous works (Johnson & Davies, 2020a, 2020b).

A little later in this same document, the following pattern emerges (Figure 37), note that the contrast has been increased to aid visualisation and that a greyscale version is available in the Appendix.
Here, it is clear that a large block of text has been created in a single edit, as almost all of the section bears the same colour code (and hence the belongs to the same rsidR value, or edit run). The single text blocks in a darker shade are where separate edits appear (though again completed in a single session). Inspection of the original text shows us that these single word edits are amendments from US spellings to British spelling. This could be considered a flag for cut-and-paste plagiarism, or contract cheating, as it would be reasonable to expect that the author would usually write in the correct language in the first instance, only needing to switch from US spelling to British spelling if the work was originally written in a different language setting.

In settings.xml there is a marker for the default language of the document. The element `<w:themeFontLang>` provides this value, through the attribute `w:val`. In the case of a document with British spelling as default, the whole element would appear as:

```
<w:themeFontLang w:val="en-GB"/>
```

If the student is using University-issued software, it could be expected (though not guaranteed) that the selected language / spelling for a document will correspond with the institution where the student is studying. Also, if the student knows they will need to submit in UK English, it could be expected (though not guaranteed) that the student will write in UK English. Of course, this may not be the case for students who have English as a second language, as they may prefer to write in their first language, and then translate into English afterwards.

In other tests on student submissions, patterns suggesting very limited editing were found. Here is an extract from a Level 4 student, who scored a 49% text match with another student identified via Turnitin (L4002) (Figure 38), note that the contrast has been increased to aid visualisation and a greyscale version is available in the Appendix.
5.8.2 Findings from initial testing

Initial results on the student files tested yield some interesting findings against the artefacts examined. A full table of results can be found in Appendix D: Table of forensic artefacts found in student submissions. Highlights are summarised in this section.

Figure 39 shows the number of words in each document (noting the challenges in getting an accurate word count discussed in 5.6.4) against the number of revisions recorded in the core.xml file. It would be expected that in general, the longer the document, the more times it is likely to have been saved, as an author edits and resaves the updated version as the document grows. For documents L4003 to L7003 this is approximately true, however, the first document (L4002) shows the greatest number of revisions (256 in total) despite having one of the lower word counts, whilst the longest document shows very few edits in relation to word count.

These values are shown as proportions in Table 20.
Table 20: Percentage of revisions in relation to word count

<table>
<thead>
<tr>
<th>Ref</th>
<th>Word count</th>
<th>Revisions</th>
<th>% revisions to word count</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4002</td>
<td>2186</td>
<td>256</td>
<td>11.71%</td>
</tr>
<tr>
<td>L4003</td>
<td>2524</td>
<td>26</td>
<td>1.03%</td>
</tr>
<tr>
<td>L4004</td>
<td>2176</td>
<td>14</td>
<td>0.64%</td>
</tr>
<tr>
<td>L4005</td>
<td>417</td>
<td>2</td>
<td>0.48%</td>
</tr>
<tr>
<td>L4008</td>
<td>3411</td>
<td>1</td>
<td>0.03%</td>
</tr>
<tr>
<td>L5001</td>
<td>2970</td>
<td>10</td>
<td>0.34%</td>
</tr>
<tr>
<td>L5002</td>
<td>1558</td>
<td>5</td>
<td>0.32%</td>
</tr>
<tr>
<td>L7001</td>
<td>1190</td>
<td>4</td>
<td>0.34%</td>
</tr>
<tr>
<td>L7003</td>
<td>3600</td>
<td>60</td>
<td>1.67%</td>
</tr>
<tr>
<td>L7004</td>
<td>18706</td>
<td>27</td>
<td>0.14%</td>
</tr>
</tbody>
</table>

The relatively high proportion of revisions to word count for L4002 is interesting. The TotalTime figure (see Figure 40), which represents the amount of time the document was open for editing, is around 36 hours. All the documents sampled (excluding L4005 which showed a zero TotalTime value) range from between 1.5 hours to 8 days of editing, but none have such a high revision count. This value therefore suggests that extensive editing was carried out on this document, despite there being a fairly average number of rsidR revision identifiers within the document.

5.8.3 Values from app.xml

Next the total editing time was analysed. According to Microsoft, ‘TotalTime’, which can be found in app.xml, represents the number of minutes that the document has been open for editing since it was created (Microsoft.com, 2010). This differs to the other time noted as part of the experiments, which is that of the time difference between when the document was first created and when it was last modified (calculated from core.xml). Indeed, a creation date that precedes the assignment release date would be suspicious (indicating that perhaps work from a previous assignment period has been used as a starting point).

Didkriksen (2014) suggests that “If the recorded editing time shows an unrealistic low number of minutes compared to the amount of content in the document, this could be an
indication that the document was composed based on other sources, which could be interesting in forensic investigations involving theft of intellectual property”.

![Figure 40: TotalTime (editing time) in relation to word count](image)

Figure 40 shows the number of words represented against the total amount of time (in minutes) spent editing. Again, it could be expected that the amount of time spent editing would increase proportionately with the number of words, but this is not clearly reflected in the above data analysis. Indeed, the longest document (L7004, comprising close to 20,000 words) has one of the lowest editing times (excluding L4005 which shows an editing time of zero). It could be argued that a longer document may be built by creating a number of smaller documents as chapters and then compiling these into a final document, and this could explain the low editing time.

Next, the number of total rsidR values was compared against word count. The results can be seen in Figure 41. rsidR values represent editing sessions, which are defined as a single session where edits / additions are made on a document.
Here, there is a general correlation of the number of rsidR values to the word count of each document. It is interesting that the number of rsidR values in L7004 is high (corresponding to the higher word count), but that both the TotalTime (total edit time) and the number of revisions is very low.

In contrast to this is the number of unique rsidR values in relation to word count (Figure 42), where L7004 shows very few unique rsidR values. This means that whilst a lot of edits were carried out, there were all done in a small number of separate sessions. This could indicate that the work was copied into the document in chunks and quick, minor edits were carried out.
throughout. It should also be noted that it may be because this document was written in chunks in separate documents and later compiled into a single submission:

![Figure 42: Ratio of words to unique rsidR values](image)

Table 21 shows the ratio of words against total rsidR values as well as against unique rsidR values. This shows the average number of words that shares a common rsidR values (arbitrarily calculated using word count / rsidR total). Some further testing was done on genuine / authentic student submissions but it would be useful to build a more reliable picture of what is ‘normal’ for authentic samples. The more words per edit run (rsidR), the more likely it is that work has been copied from elsewhere (as it is difficult to write many words without needing to correct errors or to make minor edits). As can be seen from this table, the number of words per unique rsidR value in L7004 implies that the student wrote on average 1558 words in each editing session – a truly unlikely accomplishment.
L5001 was a particularly interesting document to analyse. This was a self-reflective assessment, where the student was required to reflect on an activity carried out, referencing known theories during this process. The document highlighted an unusually long edit time of over 61 days. The assessment was set at the very start of the academic year (Sept 20) with a submission date in December, so it is feasible that the student opened a document to start work when the assignment was released and continued to work on the same document throughout the duration of the topic. Also evident in the findings are a large number of \(<w:14 textID 77777777>\), these are flags which have been found appearing frequently in text that is copied and pasted from other sources. In addition, there are 341 \(<w:rFont>\) elements, indicating that a significant amount of text (given this is an assignment of less than 3000 words) has needed to be reformatted and yet the core.xml file records only 10 revisions in total. Since this was a reflective piece of work, it should be less likely to have been copied, but these findings indicate it would be worth reviewing the document manually and perhaps discussing the content with the student in question.

L4009 was also an interesting sample. This document yielded no rsidR values at all. Whilst the \(<w:r>\) tag was present, and results showed that there were over 225 runs, no edit values were found in the XML file. This result was verified by manually unpacking the file to ensure

<table>
<thead>
<tr>
<th>Ref</th>
<th>Word count</th>
<th>Ratio of words to total rsidR</th>
<th>Ratio of words to unique rsidR</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4002</td>
<td>2186</td>
<td>11.63</td>
<td>27.67</td>
</tr>
<tr>
<td>L4003</td>
<td>2524</td>
<td>8.09</td>
<td>37.12</td>
</tr>
<tr>
<td>L4004</td>
<td>2176</td>
<td>10.07</td>
<td>43.52</td>
</tr>
<tr>
<td>L4005</td>
<td>417</td>
<td>3.28</td>
<td>14.38</td>
</tr>
<tr>
<td>L4008</td>
<td>3411</td>
<td>15.79</td>
<td>155.05</td>
</tr>
<tr>
<td>L5001</td>
<td>2970</td>
<td>12.53</td>
<td>99.00</td>
</tr>
<tr>
<td>L5002</td>
<td>1558</td>
<td>26.41</td>
<td>129.83</td>
</tr>
<tr>
<td>L7001</td>
<td>1190</td>
<td>7.26</td>
<td>108.18</td>
</tr>
<tr>
<td>L7003</td>
<td>3600</td>
<td>13.90</td>
<td>52.17</td>
</tr>
<tr>
<td>L7004</td>
<td>18706</td>
<td>24.61</td>
<td>1558.83</td>
</tr>
</tbody>
</table>
that there were no values present and that it was not an error in ‘Clarify’. Other elements were present, such as <w:sz> and <w:shd>. In addition, there are many font elements present. Also missing are core.xml, app.xml and settings.xml. This suggests that it was a document that was created in another software package. It is unlikely to be a Google doc, as this typically leaves at least a single rsidR value in the document.xml file. However, it could be a Pages document, or Libre Office and it would be useful to establish what artefacts could indicate these alternative submission formats.

5.8.4 Visual representations

The data from these investigations provided good evidence, albeit inconclusive. The visual representations, however, were very compelling. Whilst it can be difficult to quantify at what point a value should be considered indicative of misconduct, looking at the text with colour coding, or by way of a barcode can help.

In Figure 43, it is easy to see how the work of L5002 appears chunked by edit runs. There is no evidence of revisiting work once added to the document (as in minor edits and corrections). This bears the hallmarks of sentences and phrases copied and pasted into the document in a patchwriting type process. A greyscale version is provided in the Appendix.
The following work is an in-class test (Figure 44). The edit pattern looks less suspicious here:

![Figure 44: Visual representation of L4005, in class test](image)

The following example (L4004) (Figure 45) shows significantly more editing and the colour coding is less indicative of cut-and-paste plagiarism.

![Figure 45: Visual representation of L4004, where a sequence of editing can be seen](image)

In the following document, L7004, there are long swathes of un-edited text (Figure 46). Given that this is a 20,000-word Level 7 assignment, it is possible that these chunks of text represent entire chapters, perhaps copied from a separate document into the final submission (greyscale version in Appendix). This highlights the importance of reviewing the original work alongside the analysis from ‘Clarify’ in order to ascertain why such patterns of editing could appear.
5.8.5 Limitations

One of the major limitations at this stage of the research was gaining access to student files in their original Word format. At the institution in question, files relating to academic misconduct cases were stored in a shared area for the student casework team. In many cases, the student submission was stored as the PDF output from Turnitin, as opposed to the original file. This required then accessing the original files via the virtual learning environment (VLE) which was a lengthy process. It also highlighted the issue that without the original file, it is not possible to carry out these forensic analyses of work submitted, and begs the question as to whether assessors should insist on students submitting their work in the original format. Some students submit in PDF format, so even with access to the relevant VLE it would not be possible to review the document forensically to the same extent. In some cases, where the original Word file had been stored in the student casework area, it had been opened and reviewed by the assessor, the academic reporting the case, or the student casework officer, which meant that some of the metadata was altered (such as the Last Modified user details).

A recommendation coming from this stage of the research is therefore that academics should insist files are uploaded to the VLE or text-matching software in their native Word format, and that if misconduct is suspected, a duplicate of the file is created in File Manager before any detailed analysis is carried out to ensure an original version still exists. This would then follow digital forensics chain of custody guidelines to maintain the evidential integrity of the file.
5.9 Summary of Chapter 5

This chapter explained the initial development of an automated tool for detecting academic misconduct, whether that be copy-and-paste plagiarism or contract cheating. The chapter began by explaining the structure of a Word document and how this is relevant, followed by a discussion of the essay writing process and how this affects the underlying metadata. A contracted assignment was reviewed as a proof of concept for using digital forensics methods to explore the document submitted by the student and whether such analysis can provide any useful forensic artefacts that indicate the authenticity of a document. Following a successful proof of concept on this contract cheating case, investigations were carried out on a series of documents to see whether it was also possible to detect copy-and-paste plagiarism, and if so, what sort of information could be gathered from the unpacked Word document. Again, this work showed promise, and following this, iterative investigations were carried out to understand the editing process and resulting forensic artefacts in more detail and to establish whether this information could potentially be useful in detecting academic misconduct, during which stages ‘Clarify’ was developed.

Following these activities, an exploration of the other files packaged within the Word document structure was carried out, to determine whether these may also provide interesting information about how the document was created. Finally, the reporting mechanism, a web-based interface, for the findings was tidied up and made more user-friendly to ensure that the information within it could be used for an academic misconduct panel if academic misconduct was suspected. The prototype of ‘Clarify’ was completed at this point, and the next chapter provides a detailed explanation of the programme outputs along with examples to highlight its key features and discussions as to what these features show in regard to the authenticity of the file in question.
Chapter 6: Discussion and Final Working Prototype of ‘Clarify’

The findings from the initial phase into student perceptions around plagiarism and contract cheating were not surprising. Most students knew what plagiarism was, but very few claimed to know what academic integrity was. Many students knew of a fellow student who had cheated in some way or another. Their understanding was very much around detection and punishment, rather than harbouring proactive, positive views on the value of academic integrity. Hearing the student voice in relation to academic integrity, and their understanding (or otherwise) thereof plays an important role in prevention of academic misconduct, as this can help institutions target their narrative around academic integrity in the most appropriate way for their students.

An important process in the development of any student lies in the essay writing process. Writing is a skill and the process of reading, digesting and then referencing previous research, followed by critical discussion and then creation of new knowledge is integral to learning. This process can be exploited to help assessors determine the authenticity of the work they are reading, because the essay writing process itself creates some interesting digital forensic artefacts although these are rarely considered when exploring academic misconduct detection and many assessors may not even be aware of their existence. Discussed in greater detail during the Investigations phase of the research (see section 5.2), writing is a process of evolution. Whilst all writers will have their own methods, most will start with an idea, theme or question. They will note down ideas, refer to existing sources and begin to document their own thoughts. They will go back and revise these words, possibly many times over, adding, amending and deleting text until they are content with it. Technology helps greatly in this process, as software allows authors to build a document gradually over time, with endless editing possibilities, and this process can be displayed in a visual way using ‘Clarify’. By exploring the different XML files within a Word document, it is possible to extract much more data than from a single submitted piece of work.

Current methods of detecting academic misconduct focus almost exclusively on the textual, or intrinsic, content of the document submitted. This is primarily through text-matching software, where the document is checked against online text sources or offline repositories for similarity, and a ‘match’ percentage is returned. This works well as an indicator of copy-and-paste plagiarism, but does not help if the work is original but not written by the submitting student. The other growing method of detecting misconduct is through ‘authorship
tools’. These dig a little deeper into the text by analysing the style of text that is written using language, n-grams, punctuation, phrasing and so on, and comparing the styles from one document to another. If a student submits a document that is stylistically very different to the other work they have submitted during their programme of study, then this will be flagged. If several students submit work that is stylistically very similar, then this will also be flagged.

The work of Foltýnek et al. (2019) and Chowdhury and Bhattacharyya (2018) described in the Literature Review provides an extensive list of plagiarism detection tools, but of the 37 tools described, only five use intrinsic methods of detection (stylistics and linguistics). The remainder use extrinsic methods (text-matching against external sources) as discussed in the literature review.

What none of these existing tools and techniques examine is the underlying build of the document. Decompressing a Word file back to its component parts as shown by ‘Clarify’ provides the investigator with detail that is otherwise impossible to see, such as patterns of editing, font and formatting changes, uncropped image files, software owner, editing time, number of revisions and other data that can help build a picture of how the document was constructed, and from there whether it shows signs of plagiarism, contract cheating or collusion.

In Chapter 5 a range of investigations was carried out in order to develop the functions and accuracy of the plagiarism detection tool, ‘Clarify’. Results are shown by way of an on screen report making it simpler for the investigator to review the findings, and see the key features of the document being analysed. In chapter 6, a full overview of the tool is provided and the outputs discussed. This includes information about the software environment required for the tool to be run in, how files should be provided for analysis, and what sort of information results from the analysis. Justification for choices made about outputs and their relevance are included here, along with a selection of sample files that demonstrate some of the key features as well as an example of a report summary showing what information ‘Clarify’ could provide.

6.1 Setting up the ‘Clarify' environment

The software is written in PHP, a programming language that is both popular and open source. It was chosen because it works well as a web interface and pre-processes scripts on the server before presenting the results as HTML in a web page. The software version used
for this research is PHP 7.3, programmed in JetBrains’ PHPStorm. In order to run the software, a server set up is required. For this research, MAMP 6.2 was used on an Apple Mac computer. Again this is open source software, meaning that anyone wishing to test ‘Clarify’ can set up their own server environment, matching the researcher’s environment if they wish and run the software from a single computer. Both the index and PHP file must be stored within the server folder, under the ‘htdocs’ folder. All the corresponding files which are to be analysed must also sit within this folder, in a subfolder called ‘files’. A screenshot of the MAMP folder structure is shown in Figure 47.

![MAMP Folder Structure](image)

Figure 47: Clarify folder structure

An index file was created to list any files that are available for analysis, and this then calls the selected file at the same time the PHP code is executed, meaning that the PHP code analyses only the requested file. The user can then return to the index to request further files as they wish. For files to be visible within the index, they must have both a docx version (the original of the student submission) and a PDF version. At the time of writing, an automated way of generating the PDF has not been added, instead, the investigator is asked to provide both the original docx file and a PDF. The safest way to do this to ensure that no edits are inadvertently stored on the original word file is to duplicate the file in File Manager first, and then open the duplicated file and save it as a PDF by using ‘File’, ‘Save as’ and then selecting PDF as the file type. The name of the PDF can then be changed using ‘Rename’ to exactly the same as the original word file.
Once the environment is set up, the index can be loaded into a web browser. This simply provides a list of files available for analysis. A file will appear only if it has a PDF version stored with the same name in the same folder.

The user selects the file they wish to examine, and this is then pre-processed and the results presented in the ‘Clarify’ screen.
The web page opens with a note to check the file being used, along with the associated PDF. The code version is included along with the name of the file being analysed and the document type.

6.2 PDF

The PDF view of the document opens in a frame on the page, to facilitate checking of the code. This is deemed important because it makes the analysis of the document easier to follow. Where a word sits inside an edit block for example, the investigator can look back at the PDF to see how that looks within the document. Similarly, this is important for images, as it is very easy to check how images look and confirm that they have been cropped. This information is provided in the analysis, but it can be useful to have a visual check too. The PDF is also useful as tables, lists of content and reference lists can have quite extensive mark up, and it is easy for the assessor to check whether these exist within the document and may have contributed to some more unusual results.

6.3 ‘Clarify’ results section

Following the PDF output the results of the analysis are presented. Sections are broken down and hidden using CSS div tags that can be toggled on and off (hidden or visible). Various files are analysed during the processing, including document.xml, core.xml, app.xml, styles.xml, settings.xml and _rels/document.xml. These provide the most useful digital artefacts. Each of these files can be displayed in their XML code format on the webpage using the toggle div. This is useful for checking and cross-referencing the results, and was particularly useful during debugging for confirmation that the results were accurate.

6.4 Document.xml output

Within this section it is possible to render the text content of the document to the screen. Each word is coloured according to the rsidR (edit run) it was created in, by converting the rsidR value to its hexadecimal colour value.
A long passage of text in one colour means that all the text was written at the same time (no saves or edits during its creation), whereas in the example above (Figure 50) which is an authentic piece of work, text is shown in many colours, representative of the editing that has been carried out. Where a long passage of text appears in the same colour this could indicate that the text has been copied (it is very difficult to write a long passage of text without making an error or needing to go back and edit it as some point). The text blocks are broken up with a red pipe (|). Text blocks (w:t) sit inside a paragraph tag (w:p) and / or run tag (w:r) but text runs may also be broken up by spelling errors, grammatical errors or formatting features (e.g. superscript text, font changes, tabs etc). For an example of how a text run is broken by a language / spelling mismatch see Sample Figure 61 to Figure 65.

In addition to viewing the text in this way, colour blocks are also a viewing option. When selected, the text can be viewed alone, alongside a colour block that represents the edit value (rsidR) of the run. This can provide very visual feedback on how the text has been constructed – often more compelling than seeing the text itself. For an example of how the visual colour blocks work see Figure 80 and Figure 81.

**6.5 rsidR values**

The next section on the page enables the investigator to view all the rsidR values associated with the final version of the document. Note that this does not include all the rsidR values in the unzipped version, as there are likely to be rsidR values that have been deleted from the
document (see 5.7.3). Firstly, a chronological list of rsidR values is provided, showing all the rsidR values still remaining in the document.xml file in the order in which they appear, and with duplicates where they appear. Files with rsidR values that are duplicated many times may indicate that a lot of edits were done at the same time – these would have to be done very close together in time in order to share the same value, and could indicate someone sweeping through a document to make the same change over and over (e.g. an institution name).

Figure 52: Clarify output – extract of rsidR values from 10-volunteer_authentic1a.docx

Figure 53 shows the list of rsidR values appearing in an authentic student submission, in the order that they appear in the document. There are a number of duplicate values, particularly 00851206, which shows that a series of multiple edits were carried out in close proximity to each other – perhaps a table of contents and corresponding page numbers for example, but there are also many different rsidR values present, indicating lots of editing sessions overall, which is to be expected from authentic student work.

The next option (Figure 53) provides list of the unique rsidR values (regardless of how many times they appear within the document). Alongside each value is the number of times it appears in the document – useful for checking as indicated above. In this section, the rsidR is also clickable, and when clicked the value is highlighted in yellow and the text in the previous section is shaded to show where that rsidR value appears. Again, this is useful for checking what has caused the repetition of the rsidR value.
Cropped images often appear in student submissions. Students tend to drop the entire image into the document, and then use Word’s cropping tools to remove the parts they do not want. Word stores the original image before it has been cropped, and the information in the uncropped parts of the image can provide useful information.

The results firstly indicates whether any images are present in the file:

**Images information**

If this section is empty there are no images found

**Image ID: rId6**

This instance of image rId6 has been cropped as indicated below:

- Cropped on the top by approx 60.093%

Images are simply named image1.png, image2.jpg etc and the original filename of the image is not stored. Instead, Word stores each image with an associated ID, putting the original image file into a media folder. If the image is used more than once, Word uses the same ID and image name for all instances. The XML code is marked-up in document.xml to specify how the image is displayed including any settings for cropping the image. ‘ Clarify’ reports whether any images have been found, what the image IDs are, whether they have been cropped or not, and if they have, which edge of the image has been cropped (top, right, bottom or left) and by how much. In Figure 54, the file contains an image with an ID rId6, which has been cropped on the top by about 60% of its original size (i.e. 40% of the original height of the image still remains in the submitted document file).
6.7 Barcode representation of text

Given the value of knowing where large runs of text have been added to the document with no editing (indicative of pasting text into the document), the next section extracts the word length of each text run and returns it to the page. The total number of words is calculated at this point as this appears more accurate that the Word Count provided in the app.xml file (which includes hyphens, numbers and other symbols when counting). See Figure 55.

Figure 55: Clarify output – number of words per text run represented as bar code

In addition, the number of text runs with a single word in is reported to the page. These may indicate spelling or grammatical errors, punctuation or potentially words that have been altered in a large chunk of text – for example to change a name. These are worth checking in the main document text to see what has caused them.

6.8 Findings summary for document.xml

A summary of the findings from document.xml is then presented in table form to aid the assessor / investigator in deciding which, if any, features may need further investigation.

6.9 Information extracted other files

core.xml

This section displays a table of the results extracted from core.xml which includes:

- identity of the person creating the file
- identity of the person who last modified the file
• the date and time the file was created
• the date and time the file was last modified
• the calculated difference between the time the file was created and last modified
• the number of revisions recorded by Word

Information extracted from core.xml

<table>
<thead>
<tr>
<th>Category</th>
<th>Results</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created by:</td>
<td></td>
<td>An unrecognised author should be queried</td>
</tr>
<tr>
<td>Last Modified by:</td>
<td></td>
<td>Unrecognised last modified by author should be queried</td>
</tr>
<tr>
<td>Date created:</td>
<td>2019-01-14 at 10:28:00</td>
<td>A ‘date created’ value that precedes the assessment release date should be queried</td>
</tr>
<tr>
<td>Date last modified:</td>
<td>2019-01-31 at 12:40:00</td>
<td></td>
</tr>
<tr>
<td>Difference between creation time and last modification:</td>
<td>0:0:17:2:12 (Years:Months:Days:Hours:Minutes)</td>
<td></td>
</tr>
<tr>
<td>Number of revisions:</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>

Figure 56: Clarify output – core.xml findings

As noted previously, these values can be manually altered and are not fully reliable.

app.xml

This section reports on information such as the version of Word being used, the total time spent editing (TotalTime), the approximate number of words in the document, and it summaries some of the key findings that are of interest to the investigator, calculating the radio of words to runs and the ratio of words to total rsidR values.
**Information extracted from app.xml**

<table>
<thead>
<tr>
<th>Category</th>
<th>Results</th>
<th>Flag score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word version:</td>
<td>16.0000</td>
<td></td>
</tr>
<tr>
<td>TotalTime* is:</td>
<td>1187 minutes</td>
<td></td>
</tr>
<tr>
<td>Number of words given in app.xml:</td>
<td>2200</td>
<td></td>
</tr>
<tr>
<td>Number of words via text count (excludes numbers):</td>
<td>2413</td>
<td></td>
</tr>
<tr>
<td>The document is shared:</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>Run count:</td>
<td>323</td>
<td></td>
</tr>
<tr>
<td>Ratio of words* to runs (w:r):</td>
<td>7.47 words per edit</td>
<td></td>
</tr>
<tr>
<td>Radio of words* to total rsidR values (w:rsidR):</td>
<td>7.91 words per rsidR</td>
<td>0</td>
</tr>
</tbody>
</table>

*TotalTime represents the number of minutes that the document has been open for editing since it was created

**words value is taken via calculated count on text content as this is more accurate than app.xml value

Figure 57: Clarify output – app.xml findings

**settings.xml**

This section provides the default language setting for the document. This can be useful information, if, for example, there are edits throughout the file changing US spelling to English (UK) spelling, but the default language of the document is already English (UK) and would indicate that the text has been copied from elsewhere. A full list of the rsidR values that have appeared in the document at any time is listed here. If there are many more rsidR values in settings.xml than in the document.xml file, it suggests significant deletions have occurred, as all rsidR values that have ever appeared in the document are retained within the settings.xml file, but if the corresponding content is deleted from the document, then the rsidR tag that marked it is removed from document.xml. The report therefore feeds back the difference between the number of rsidR values appearing in the settings.xml file in comparison to the number appearing in document.xml.
Note that in the Word application, the language setting is described as ‘English (United Kingdom), and that the equivalent UK spelling is recorded as en-GB in the XML file (or en-US for American English, for example).

styles.xml

This file is less useful that the others for the purposes of an investigation. However, it contains the relationship between the images and the ImageIDs which is needed for exploring the cropped image files.

6.10 Uncropped image files

This section displays to screen all images which are in the document in their full, uncropped form. The report also states whether the image has been cropped or not, and if so, which part has been cropped (left - l, bottom - b, top - t, right - r). In future work it would be useful shade out the part of the image that has been cropped to highlight to the investigator which parts have been removed. The cropped parts can provide useful insight into where the image came from, if it is not clear from the document.xml file. In Figure 59 the reason for cropping is clear – there is a considerable about of white space that needed to be removed from the image in order to prevent a lot of white space on the page.
6.11 Flag system

During the development of ‘Clarify’ the addition of a flag system was considered. In text matching software, a final percentage ‘score’ is provided to the assessor to indicate how much of the document matches text from elsewhere. Flags in ‘Clarify’ were considered that could contribute to a final score to facilitate the swift analysis of results, where documents scoring above a certain threshold could then be deemed in need of manual review. Values such as the number of specific tags, unique rsidR values, appearances of white text, number of words to rsidR values, long text runs and so on were all tested as potential contributors to this system. However, without a large corpus of authentic work to develop benchmarks for authentic work, any such flags could be in danger of false positives, and so this was not included at this point in the software’s development.

6.12 Sample files

A number of example files have been included in the final iteration of the software. These highlight single significant aspects of the tool to demonstrate how it works. For example, there are files that have been fabricated to show white text between words, files which
deliberately use copy-and-paste, files which show US to UK spelling amendments, files with cropped images, files with very little editing and so on. There are also some authentic files, which were submitted by students who volunteered to have their work used as example test files for the software development.

The following examples demonstrate how the tool works and the results output from a variety of test cases, along with discussions as to their relevance.

6.12.1 Spelling and language – Samples 1a, 1b, 1c and 1d

A file was created with the US spelling of the word behaviour (behavior), on Word software set to English (United Kingdom) language (which would appear in the XML as en-GB).

Here is the text shown with the colour coding as part of the output:

This document uses English (United Kingdom) spelling. The word | behavior | has been added to show the xml mark up around the spelling. |

Figure 60: Clarify output - colour visualisation of spelling error Sample 1a

The colour coding indicates that all the text was created in one editing session, but the red pipes around the word ‘behavior’ indicate that some other markup is embedded here. In this case, it is because the word is spelled with the US spelling, and the language of the file is set to English (United Kingdom). The mark up therefore tells Word to display the word ‘behavior’ with a red wavy line underneath it to alert the author that there is a spelling error, using the mark-up <w:proofErr> to indicate the issue.

Displaying this text as colour blocks looks like this:
In the second version of this file, an additional introductory sentence was added, and the word ‘behaviour’ was corrected by the author to the correct UK spelling. This is how the amended text appears in ‘Clarify’:

Figure 63: Clarify output – spelling error with amendment Sample 1b

There are now two different colours in the document – the original green colour, which is the text from the first save of the file, and the orange/pink colour which shows the additions.

Figure 64: Clarify output – colour block visualisation of spelling amended sample 1b

Where the word ‘behaviour’ was updated (in this case using right click and replacing the entire word) the word now appears in the new orange/red colour to show that it was corrected at the same time as the first (new) sentence was added.

A further example on this file shows how adding just a ‘u’ to the word behavior in the first sentence results in a third colour, or rsidR value, this time just around the ‘u’ itself as only this letter was added.

Figure 65: Clarify output – visualisation of amending the spelling manually sample 1c

Finally, later in the results, in the Settings section, the default language for the document is extracted. In this case it is UK English, indicated by the en-GB extracted from the settings.xml file:
Edits that relate specifically to language settings can suggest that text has been copied from elsewhere. Testing shows that if a word is corrected during the typing process – i.e. the author spots the error from the red underline as it is typed, and then amends it immediately before continuing on with the rest of the typing, no mark-up is added. Therefore, it is most likely that work containing subsequent spelling mark-up has come from a source with a different default language, and has required spell checking after the text has been copied in.

In the above example, if the word ‘behaviour’ had been amended contemporaneously – i.e. typed incorrectly in the first instance, but immediately altered manually as Word flags it as incorrect, there is no additional mark-up on the text. This is because the correction is made in the same editing session as the original text is typed, and before Word has the opportunity to formally mark the text as a spelling error (Figure 66):

![Figure 66: Clarify output – spelling corrected as typed sample 1d](Image)

### 6.12.2 PowerPoint and image exploration

Whilst the tool is not yet optimised for PowerPoint files, a very interesting example of image plagiarism presented itself amongst the academic misconduct files available for analysis. In this case, the assessor believed that the images submitted as development boards (mood board, colour board etc) were not the student’s own, but had been downloaded from the Internet. On questioning the student, this was confirmed as being the case, although the student claimed that they had misunderstood the brief. Plagiarism was therefore confirmed by the student, making it a clear case on which to pass judgement. On running the file through ‘Clarify’, the extent of the plagiarism is very clear, and would leave no doubt in the minds of the members of the academic misconduct panel that plagiarism had occurred, should the student have denied it, providing irrefutable evidence that the work was not the student’s own. Images from the student submission shown with the uncropped image file as extracted by ‘Clarify’ are provided here:
Figure 67: Student submission of fabric ideas

Figure 68: Internet original of student submitted image showing it has been copied from Pinterest
Design development

Figure 69: Student submission, with the implicit suggestion that the work is the student's own

Figure 70: Internet original of the artwork
Fashion illustration 1

Figure 71: Student submission

Figure 72: Original image from the Internet
Fashion illustration 4

Figure 73: Student submission

Figure 74: Original Internet version
In another example, uncropped images within a student submission showed what appear to be a submission file name. This is student work submitted in the year 2022:

![Figure 75: Student submission citing the source as Pew Research Centre](image1)

![Figure 76: Uncropped version of graphic showing a filename [redacted] dated 2019](image2)
This image was therefore taken from a PDF file (not from the cited source) and shows what looks like a possible submission filename. This could potentially be cross referenced to 2019 submissions to search for an older student submission that this may have been copied from.

6.12.3 Hidden (white) text

The file random_white_space utilises a method of plagiarism obfuscation that inserts random characters between the words copied on the page in a white font. This effectively renders them invisible to the reader, unless the whole document is highlighted and the text turned to another colour (or reset as black), and shows a deliberate attempt to fool text-matching software. In this example, the document appears as in Figure 77 to the reader:

The various sub-fields of AI research are centered around particular goals and the use of particular tools. The traditional goals of AI research include reasoning, knowledge representation, planning, learning, natural language processing, perception and the ability to move and manipulate objects. General intelligence (the ability to solve an arbitrary problem) is among the field's long-term goals. To solve these problems, AI researchers have adapted and integrated a wide range of problem-solving techniques -- including search and mathematical optimization, formal logic, artificial neural networks, and methods based on statistics, probability and economics. AI also draws upon computer science, psychology, linguistics, philosophy, and many other fields.

Running this file through ‘Clarify’ shows a high appearance of white text in the report, which would warrant further investigation (as there is very little reason why any white text would appear in a Word document at all):

<table>
<thead>
<tr>
<th>Total number of appearances of white text:</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
</tr>
</tbody>
</table>

In the Document Output section, the white text is clearly visible as it is displayed in the colour of the rsidR value that represents the editing session it was created in. In this case, the hidden text is white ‘x’s’. This makes it impossible for text-matching software to pick up the copied passages of text.
Of course, white text is only hidden when it is on a white background, but even if the text had been blue on a blue background it is very clear to identify as an anomaly in this visual output, where there is a single letter text run at each instance of the letter ‘x’ and the x’s are clearly visible.

6.12.4 Copy-and-paste

The next sample takes an authentic piece of work (4a-Copy_and_Paste_book_chapter) and analyses it. The content of the document is then copied to a new document and resaved to explore the differences in the results. Key findings are given below.

In the first instance, the visual appearance of the text when colour coding is applied is very compelling:

After the document is copied and the copied version analysed using ‘Clarify’, the following colour visualisation is created:
In the copied version, there are very few rsidR values remaining, as these are stripped out during the copying process, so the visualisation is very different.

Looking at the two document.xml summary files side by side, the following information can be gathered:

<table>
<thead>
<tr>
<th>Item</th>
<th>Original document</th>
<th>Copied document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of paragraph elements</td>
<td>196</td>
<td>197</td>
</tr>
<tr>
<td>Total number of runs (w:r)</td>
<td>1340</td>
<td>382</td>
</tr>
<tr>
<td>Total number of rsidR attributes</td>
<td>1150</td>
<td>197</td>
</tr>
<tr>
<td>Total number of unique rsidR values in document.xml</td>
<td>511</td>
<td>2</td>
</tr>
<tr>
<td>Total number of rFonts</td>
<td>1066</td>
<td>331</td>
</tr>
</tbody>
</table>

This shows that the number of paragraph elements remains consistent (as the amount of text is the same in both documents), but that the number of runs is significantly less in the copied document. As a run defines a region of text with common properties, not the number of edits (which is represented by the rsidR element), the significantly reduced number of runs is likely to be due to the fact that extraneous code and edit information has been removed during the copying process. The total number of rsdiR attributes also reduces significantly in the copied document, from 1150 in the original document to just 197 in the copied document.
This demonstrates how the editing information of the document is removed during the process of copying. The number of unique rsidR values also reduces drastically, from 511 in the original document to just 2 in the copied document. This effectively means that only two editing sessions were carried out in the copied document. Note here that this document is approximately 7,500 words long, so two editing sessions to complete that amount of writing is highly suspicious. There is also a significant reduction in the total number of rFonts seen, again this is likely to be due to the loss of extraneous information in the copy and pasting process.

6.12.5 Contract Cheating Assignment 1

Reviewing the results for L006_ContractCheating yields some very interesting results. Exploring the unique rsidR attributes in the document shows that a small number of values appear extensively throughout. For example, 00B23211 appears almost 450 times within the document, meaning that 450 changes were made in a single editing session in non-consecutive parts of the document – quite a feat. Toggling this rsidR value on so that it can be seen within the text of the document shows that there are some large blocks of text that use this rsidR value, and that this is also the value given to much of the numbering throughout the document, suggesting that the numbering for the sections were added in one go, which would seem reasonable. However, looking at the XML manually reveals that there are a significant number of redundant formatting attributes that take this value. For example, following one sentence of text, this line of code appears:

<w:r w:rsidR="00B23211" w:rsidRPr="008A45E2"/>

This marks the start of a run (w:r), with an edit run value (rsidR) of 00B23211. This single line of code appears 437 times within the document. All instances of this edit value would have been created at the same editing point and yet in most cases, there is no text. Where there is text, it represents the insertion of a number (reference), space or punctuation mark. Further testing would be needed to reliably establish how this might occur, but one theory could be that in bringing the text in from elsewhere there is extensive redundant formatting – perhaps relating to paragraph spacing, or double line spacing.
6.12.6 Contract Cheating Assignment 2

File reference L4007 represents a Level 4 (1\textsuperscript{st} year Undergraduate) submission found guilty of contract cheating at an academic misconduct panel. Putting this submission through ‘Clarify’ yields the following results:

![Clarify output - submission found guilty of contract cheating](image)

The output has been considerably reduced in size but shows a 3500 word document, with large blocks of colour, meaning that much of the document was created in one go, with minor edits throughout. These minor edits are for single words throughout the document, where there appear to be word substitutions (‘significance’, ‘thus’, ‘schemes’ – black blocks, and ‘curricula’, ‘selecting’, ‘amenability’ – pink blocks), carried out in several separate editing sessions. There are also amendments throughout for citations (orange blocks), as well as minor typographical adjustments (commas, spaces and brackets). The number of unique rsidR values in the entire document is only 11, meaning that only 11 separate editing sessions were required to complete this whole essay. To summarise, this is a reasonably long document which shows a small amount of editing. The editing that has been carried out consists of minor amendments, suggesting that the content has been added the file in one or two blocks and then quickly amended. This evidence would have been a very useful additional to an academic misconduct panel. This document also includes 91 appearances of the \textless webHidden\textgreater tag. This tag means that content should not be displayed in a web page. Typically, it does not appear when creating content from scratch (i.e. typing directly into a Word document), but further testing would provide more conclusive reasons as to why this tag may appear in the XML.
Earlier in the thesis the case of contract cheating that triggered the initial concept of developing the tool ‘Clarify’ was discussed. Reviewing this document in ‘Clarify’ provided some results of note. Firstly, the rsidRoot values of the Contracted Author’s document was compared to the submission by the student. These values were different, which indicates that the student created a new blank document into which to paste the contents of the work written by the Contracted Author. The list of rsidR values appearing in the settings.xml files for both documents was then examined. In the Contracted Author’s document there are 118 unique rsidR values in settings.xml, and 88 unique rsidR values in document.xml. This is considerably different in the student submission which shows 143 unique rsidR values in settings.xml but only 15 unique rsidR values in document.xml. Word counts for the documents are similar at 3836 in the Contracted Author’s work 4074 in the student submission (see Table 22):

### Table 22: Essay Mill and Student Submission comparison

<table>
<thead>
<tr>
<th></th>
<th>Contracted Author’s document</th>
<th>Student Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word count</td>
<td>3836</td>
<td>4074</td>
</tr>
<tr>
<td>Unique rsidR values in</td>
<td>118</td>
<td>143</td>
</tr>
<tr>
<td>document.xml file</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique rsidR values in</td>
<td>88</td>
<td>15</td>
</tr>
<tr>
<td>settings.xml file</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Taking this a step further, it was possible to extract the rsidR values from the settings.xml files in both documents from the ‘Clarify’ report and compare them in an Excel spreadsheet. Of the 118 unique rsidR values in the Contracted Author’s work, only 2 do not appear in the student submission. That is, all edits that exist within the settings.xml file of the student submissions are duplicates of those within the Contracted Author’s work. The number of possible hexadecimal codes is approximately 16.7 million (16^6), and given that each rsidR is allocated randomly, the chance of matching one of those 1 in 16.7 million values in a second document is infinitesimally low. This therefore highlights that the work between the
two documents is highly similar and originates from the same source, despite the rsidRoot values being different. A further interesting artefact is uncovered by ‘Clarify’ and that is that there is an rsidR value that appears in the document.xml file but which does not exist in the settings.xml. Since all edits carried out within the document will automatically be stored in the settings.xml file as the document is created and modified, the only reason for a value appearing in only the document.xml file is that it has been copied into the file from another source (i.e. the text that it belongs to came from elsewhere). In this case the value is 00AC6614 and on searching the XML this value can be found towards the start of the document in what appears to be redundant code as it is a series of paragraph tags (<w:p>) with no content, following the table of contents. Indeed, this series of redundant paragraph tags include the value w:14:textId="77777777" and in testing and investigations this value appears to be a legacy of copied and pasted content.

6.12.8 Punctuation edits

In this example (Figure 83), the visualisation of the essay shows many minor edits that were all carried out in a couple of separate editing sessions. The purple edits are made up primarily of commas (circled), spaces, inverted commas and capitalisation of initial letters at the start of sentences. This could be considered reasonable practice – i.e. the tidying up of existing written work, but the vast amount of orange shows that the majority of text was added to the document in one go – suggesting that it is almost definitely copied from elsewhere.

Figure 83: Clarify visualisation showing punctuation edits
6.13 Example ‘Clarify’ report on a plagiarism case

By way of an example of reporting from the tool, a submission suspected of plagiarism was put through ‘Clarify’ for examination. The following details were obtained from the report:

- The document was created using software licenced to the institution that the case originated from.
- The original document was created in 2019 (and submitted by a student in 2022).
- The document was open for 19 hours editing time.
- The creator of the document was the same as the person who last modified the document in June 2022.
- There were 184 document revisions.
- There were 267 edit runs within the document, but 170 lost edit run values, suggesting significant deletion has occurred.
- There are lengthy text runs with no editing – including some with word lengths of 141, 125, 94, 84. These would be extremely difficult to type without requiring any sort of further editing.
- There are a large number of font changes (possibly due to tables, content lists, but also can be because reformatting was required to match the default font of the document), also a large number of font resizes.
- Visualisation shows that each ‘chapter’ or section of text was edited in a single run.
- Cropped images are used within the document, and these reveal another document with a filename that looks like a student submission with a date of 2019 (i.e. 2019_essaytitle_FINAL.pdf).

This data is extracted automatically through ‘Clarify’, with manual intervention only required at this stage to extract the most interesting findings into a single document. Later iterations could automate this aspect of the reporting too.

6.14 Testing ‘Clarify’ with volunteered authentic work

As previously noted, students were given the opportunity to volunteer authentic work to be analysed by ‘Clarify’. The purpose of the tool was made clear to all students contacted, but despite two attempts to several hundred students, only six samples were submitted: two each from three students. Of these samples, two were primarily computer code as required by the
assignment brief. ‘Clarify’ reports this as large chunks of code with no editing, which would correlate with code written in a software package and then imported into the Word document for submission. A third sample consisted of a gaming assignment and was also almost exclusively imported from gaming software. This leaves only three samples and whilst these were reviewed and demonstrated the expected outputs these cannot be used for statistical significance. What is worth noting is the average words per run and the ratio of words to rsidR (edit) values in comparison to the work that had been found guilty of academic misconduct:
Table 23: Comparison word ratios between authentic work and academic misconduct samples

<table>
<thead>
<tr>
<th>Sample</th>
<th>Volunteer or Academic Misconduct</th>
<th>Average words per run</th>
<th>Ratio of words to rsidR values</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1a</td>
<td>Volunteer</td>
<td>7.47</td>
<td>7.91</td>
</tr>
<tr>
<td>V1b</td>
<td>Volunteer</td>
<td>4.35</td>
<td>4.83</td>
</tr>
<tr>
<td>V2a</td>
<td>Volunteer</td>
<td>1.18</td>
<td>4.56</td>
</tr>
<tr>
<td>V2b</td>
<td>Volunteer</td>
<td>1.76</td>
<td>4.63</td>
</tr>
<tr>
<td>V3a</td>
<td>Volunteer</td>
<td>1.44</td>
<td>1.34</td>
</tr>
<tr>
<td>V3b</td>
<td>Volunteer</td>
<td>4.16</td>
<td>4</td>
</tr>
<tr>
<td>L4002</td>
<td>Academic Misconduct</td>
<td>9.18</td>
<td>11.63</td>
</tr>
<tr>
<td>L4003</td>
<td>Academic Misconduct</td>
<td>6.47</td>
<td>8.09</td>
</tr>
<tr>
<td>L4004</td>
<td>Academic Misconduct</td>
<td>9.46</td>
<td>10.07</td>
</tr>
<tr>
<td>L4005</td>
<td>Academic Misconduct</td>
<td>3.16</td>
<td>3.28</td>
</tr>
<tr>
<td>L4008</td>
<td>Academic Misconduct</td>
<td>8.86</td>
<td>15.79</td>
</tr>
<tr>
<td>L5001</td>
<td>Academic Misconduct</td>
<td>10.17</td>
<td>12.53</td>
</tr>
<tr>
<td>L5002</td>
<td>Academic Misconduct</td>
<td>12.08</td>
<td>26.41</td>
</tr>
<tr>
<td>L7001</td>
<td>Academic Misconduct</td>
<td>5.41</td>
<td>7.26</td>
</tr>
<tr>
<td>L7003</td>
<td>Academic Misconduct</td>
<td>9.57</td>
<td>13.90</td>
</tr>
<tr>
<td>L7004</td>
<td>Academic Misconduct</td>
<td>12.27</td>
<td>24.61</td>
</tr>
</tbody>
</table>

Whilst this is very small sample, it can be seen that the average words per run over the six volunteer assignments works out at 3.39 words per run, whilst for the academic misconduct samples it is 8.66 words per run. For the ratio of words to rsidR values, the volunteer assignment average is 4.55 and the academic misconduct samples average is 13.36. With copy-and-paste work, or contracted work that has had only minor edits, it could be expected that the average words per run, and in particular the ratio of words to rsidR values would be higher, and this is evidenced here, albeit with the caveat that this is a very small sample set for comparison.
6.15 Ethics discussion

A full description of the ethical considerations of carrying out this research was discussed in Chapter 2. However, some further discussion around the challenges and impact of these challenges is worth revisiting as part of the discussion. For the primary research, ethical approval was sought from the faculty ethics champion, and granted prior to the first collection point. Confirmation of ethical approval was circulated to each of the HOLSTEs when identifying which cohorts could take part. Three of the faculties were satisfied to accept the original ethical approval, but the fourth had concerns. They were of the opinion that anonymity should not be guaranteed and that if a student identified themselves as having cheated in an assessment, then they should subsequently be reported to the course Academic Subject Manager. To not do this, in their view, could bring the University into disrepute.

There were also some concerns around subjects that operated a Fitness to Practice policy, where the impact on any student found to be engaging in academic misconduct could be more significant to future careers choices and employment. However, had students not been guaranteed anonymity during the research gathering, there would be a risk that they would not feel able to talk openly about their experience of academic misconduct, cheating behaviours etc, if they knew they would be reported. After much discussion, this faculty was not included in the primary research phase of the study, for fear that it would skew the results with false reporting on the number of cheating incidents. Of course, this also means that the data is missing from the overall review of the institution, as students differ greatly from one faculty to the next, and therefore the data will inherently be skewed. Furthermore, given that the methodological approach for this research is that of a case study, anonymising the data makes it more difficult to understand the reasons why students might cheat (demographic data is useful here as an example) (Cohen et al., 2018).

In Chapter 3, it was shown that blocking certain sites known to be offering essay-writing services could be one such way in which technology can support academic integrity. Other cyber security approaches can be deployed to both prevent and deter academic misconduct, but bringing a policing approach to academic integrity is not without its challenges, both practically and morally (Lynch et al., 2021). Rettinger and Bertram Gallant (2022) remind us that the focus of educators should not be on finding technological solutions to detecting academic misconduct, but on the pedagogical aspects of improving academic integrity.
Similarly, tracking student activity using learner analytics, monitoring log ins to the VLE both using location data, and date and time stamps can bring with it a big brother feel. As Johnson et al., (2023, in print) note, the police are not allowed to stop and search someone in the UK without reasonable grounds, yet some suggest it is acceptable to monitor and track student activity where there has been no cause for concern. Cyber security type technological advances that support the prevention, detection and evidencing of academic misconduct as suggested by Dawson, 2020, Johnson et al., 2022 and Murdoch & House, 2019, elicit questions around the ethical stance that should be taken. How ethical is it to ‘police’ students in this way? How far is too far? Is tracking a student’s location to see where he or she was when logging into the VLE reasonable? Is it a step too far to plot a pattern of behaviour such as suggested by Murdoch and House (2019), who described a student who was in one location when they logged in during the evening, and close by when they logged in the following morning, but thousands of miles away when they logged in to submit their assignment in between these two sessions? Yet institutions do have a duty of care to their students, so protecting them from essay mills would seem entirely ethical. They also have a duty to ensure the awards they are giving are genuinely deserved, and again, if this means more robust detection methods are needed that would seem acceptable. One possible solution would be to ask students to give informed consent that this sort of monitoring is being carried out. This would require a strategic and policy driven approach by each institution as part of a legitimate approach to safeguarding, but there are likely to be some students who feel that this may be an infringement on their privacy.

Finally, Foucault (2020) also raises the idea of ‘binary division’ which is, perhaps, a risk associated with academic misconduct detection – pigeon holing students into ‘good / bad’, ‘rule-abiding / cheaters’ – and the associated damage this can bring. This is an important consideration when considering both the motivations of students engaging in academic misconduct, as well as when considering aspects of detection that may lead educators to make arbitrary decisions about them.

6.16 Summary of Chapter 6

Development of automated software called ‘Clarify’ has been discussed in this chapter as a potential tool for the detection of academic misconduct. The key features of the software were provided, including what certain features may indicate when found in a document. The chapter also provided a detailed walkthrough of the final working prototype of ‘Clarify’. An
explanation of the environment set up was given, including software and hardware requirements, file structure and file formats. Screen shots of the output were provided along with how the software functions and what information it provides to the investigator. Finally, a series of sample files were provided to allow discussion around how ‘Clarify’ might provide new approaches to plagiarism and contract cheating detection.

As with all plagiarism detection methods, there are limitations to the tool. Firstly, the tool is only optimised for Microsoft Word documents at present. PowerPoint files have also been tested, but this is in its early stages. In addition, a manual approach to providing the PDF for the reporting is used, as automating this process was somewhat troublesome. The need for the PDF could be removed from the tool to avoid this, but currently the software checks for the presence of the PDF before including the file in the index. Secondly, just as is the case with text-matching software, any findings that result from the analysis need to be reviewed by someone familiar with the system and the assignment rubric. As mentioned throughout, there may be very good reasons why a student submission shows an unusually low number of edits – perhaps because they chose to copy the text from one file to another for some reason, to combine chapters for example, or perhaps because they wrote the work in their first language and then needed to translate it using an online translation system before copying it into the final document for submission. These cases will all need to be reviewed by someone to decide whether or not these are feasible reasons, and may require discussions or a viva with the student in question. Some evidence will be irrefutable – such as images which, when uncropped, reveal information about the source which does not tally with the reported document author, or hidden text used to confuse text-matching software. In all cases though, the tool automates the technical aspects of document investigation and evidence gathering – skills which could be expected of a digital forensics investigator, but that would be challenging and time consuming for an academic. The resulting findings are provided back to the assessor as a report that could be used as evidence in an academic misconduct panel.
Chapter 7: Conclusions and Future Work

Detecting academic misconduct when it occurs is an important part of the assessment process. As well as identifying students who try to gain credit where it is not deserved, effective detection tools can act as a deterrent to those considering engaging in cheating behaviours and can protect students from making poor decisions at a time when they feel under pressure to deliver. Current detection methods focus almost exclusively on the textual content of a document by checking for text matches with existing sources, or by using authorship tools to verify the author of the work with a degree of certainty. Whilst text matching is reasonably successful for copy-and-paste plagiarism, success in detecting work that has been contracted out still has some way to go. The recent advances in artificial intelligence and growing awareness around these tools as mechanisms for producing complex essay answers mean that new approaches to detection of academic misconduct are essential.

The research conducted throughout this PhD study provides a novel contribution to the field of academic integrity by repurposing digital forensics techniques to extract and analyse the editing metadata of a Word document. This metadata provides useful information about the way the document has been constructed and can highlight where the document does not appear to have been written in a ‘normal’ or authentic way. Academic misconduct evolves as technology evolves, and methods of detection similarly need to keep pace with these advances. The prototype detection tool, ‘Clarify’, developed during this research, has already seen much interest. Academic integrity experts from the UK, Germany, Canada and Australia have shown an interest in the tool’s development.

The research takes a holistic approach to academic integrity, because whilst educators may wish to prevent academic misconduct from occurring in the first place, the reality is that this is not possible, and so to ensure the quality of qualification, detection is needed. Two key themes therefore underpin the basis of this research, firstly that of students’ perceptions of academic integrity, and secondly the challenges of detection when academic integrity is not upheld and academic misconduct occurs.

The first phase of the research was to explore existing literature relating to students’ perceptions of academic integrity, which included developing a taxonomy of existing research in this area. This was followed with primary research to gather the views of students at the researcher’s institution. A series of detailed investigations into the OOXML format was then carried out during the second phase of the research, and alongside this was the
development of the academic misconduct detection tool, ‘Clarify’. In Chapter 6 a discussion around the findings of the two phases of the research was provided, along with a full working prototype of ‘Clarify’, complete with worked examples of the tool in use.

7.1 Key findings from ‘Clarify’

‘Clarify’ is able to automatically extract a number of digital forensic artefacts from Microsoft Word documents. Just like ‘Indicators of Compromise’ in cyber security, these artefacts can suggest that something is not usual about the way the document has been constructed.

The main features extracted by ‘Clarify’ are listed in the table below, along with a summary of what they may indicate or why they are useful:

Table 24: Summary of features provided by ‘Clarify’

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>All rsidR edits values are listed</td>
<td>Shows which content was added at the same time (i.e., within a single save point, whether user-invoked or automatic). Useful for linking edits together.</td>
</tr>
<tr>
<td>Unique rsidR values listed, along with number of times each one appears</td>
<td>An rsidR value that is used a large number of times indicates that the author has swept through the document and made a lot of changes in a small amount of time. This could be changing a specific word to a different word (e.g., Wales to England) or amending spellings (e.g., US to UK) for example.</td>
</tr>
<tr>
<td>Colour visualisations of the editing</td>
<td>Large chunks of text in a single colour mean the text was added in one go. This provides very compelling evidence of whether the document is authentic or not. An authentic document will be littered with lots of different colour throughout.</td>
</tr>
<tr>
<td>Bar code visualisation of text runs</td>
<td>Blocks of text without any breaks (e.g., edits or spelling errors) appear in a coloured block. Longer blocks indicate longer runs of unbroken text. It is difficult to create long runs of text when typing, so large blocks (e.g., more than 100 words) are suggestive of copied text.</td>
</tr>
<tr>
<td>Extraction of uncropped images</td>
<td>Displaying uncropped images from the document can reveal useful information about the origins of the image</td>
</tr>
</tbody>
</table>
| Hidden text | Hidden text is used as a deliberate attempt to fool text matching software. Currently ‘Clarify’ only flags white
text, but it also displays all text within the document so unusual characters will be displayed regardless of colour (useful if the background is not white, and text has been made to match the colour of the background).

<table>
<thead>
<tr>
<th>Appearance of textId=&quot;77777777&quot;</th>
<th>This artefact appears to occur when text has been copied from elsewhere.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Font tag extraction</td>
<td>Extensive font attributes can be indicative of copy-and-paste plagiarism</td>
</tr>
<tr>
<td>Date created and last modified dates</td>
<td>Anomalies or unexpected findings should be investigated</td>
</tr>
<tr>
<td>Author and creator names</td>
<td>Anomalies or unexpected findings should be investigated</td>
</tr>
<tr>
<td>Language</td>
<td>Each document has a language setting, which results in spelling checks. Comparing rsidR values with spelling amendments can indicate unusual writing practices.</td>
</tr>
</tbody>
</table>

One of the most compelling features of the user interface for ‘Clarify’ is the colour coding and visualisation of the text. At a glance, it is very easy to see whether a document looks authentic, because it has short text runs and lots of colour, commensurate with normal typing and editing practice. Conversely, a document that has large blocks of single colour and very long text runs is instantly questionable.

### 7.2 Research questions

This thesis posed two strands of questioning relating to plagiarism and contract cheating. Initially exploring student perceptions regarding academic integrity and where the student voice is heard in relation to this, primary research was carried out to establish whether the findings from this exploration could be verified at the researcher’s institution. This formed the underpinning of understanding around what students think about plagiarism and formed a value base on which the next stage of the research focused. To the end, the research questions for this first phase were as follows:

1.a What is the focus of existing research into academic integrity / misconduct?

1.b What types of data already exist?
1.c. To what extent is the student voice heard in existing research?

1.d. How will hearing the student voice help in reducing plagiarism, and in particular contract cheating?

In order to understand student perceptions, a review of existing literature in this field was carried out. This showed that much of the data collection of student perceptions is quantitative, using surveys with tick boxes and Likert scales to establish how students view academic integrity and academic misconduct. It also showed that surveys were typically created by academics, asking questions that educators see as relevant, rather than taking a student first approach and inviting students to discuss their views more openly. This, in part is down to the problems that are perceived around obtaining ethical approval, and this is one of the major challenges in having open discussions with students who have plagiarised – it is marginally easier with students who have been found guilty of plagiarism, but almost impossible with those who have not formally been detected.

There were several surveys discussed during the Literature Review which incorporated open-ended questions, but these yielded limited data, and there was some data from focus groups, which provided valuable contributions (e.g. Ashworth et al., 1997; Bretag et al., 2014; Michalska, 2014). However, these were again limited in number, and the researcher felt that there was more to be done in this area.

Primary research was then carried out to capture students’ perceptions of academic integrity. Understanding student perceptions helps to reduce plagiarism, as hearing the student voice gives a far better understanding of students’ reasons for this form of misconduct. What came out loud and clear from the primary research carried out (which corroborates other similar research that has since been carried out) is that whilst students generally have a good understanding of what ‘plagiarism’ is, that information is very much focused on penalty and risk to successful completion of the student’s chosen course, i.e. it is punitive. In the primary research, very few students were able to accurately define academic integrity and that as their studies progressed, they were less confident in their understanding of integrity. This leads to a recommendation that much more work should be done around developing competency in academic referencing, academic writing, and understanding the reasons for and importance of academic integrity. This phase of the research answers the first set of research questions relating to student voice. What became apparent during this stage is that the reasons students engage in academic misconduct are varied, and that students have mixed understandings of
Understanding academic integrity and plagiarism in the digital age  Clare Johnson

academic integrity and academic misconduct, such that it is unlikely educators will be able to prevent academic misconduct occurring. This is why the second part of the research is so important, that of detection. As students become ever more wise to the way in which academic misconduct can be detected, and as essay mills become more widely known about as well as more commercial in their approach to engaging with students, and with the development of artificial intelligence as a means to write, new methods of detection are needed. For this reason, the focus of the research moved into plagiarism and contract cheating detection, using digital forensics approaches that are very different to the existing approaches used for automated academic misconduct detection.

The phase 2 research questions were as follows:

2.a What tools currently exist to help detect plagiarism and contract cheating?

2.b What digital forensics tools (if any) are used in other areas that could be applied to plagiarism and contract cheating?

2.c To what extent is it possible to develop digital forensics techniques to help identify plagiarism and contract cheating?

The tools that currently exist to help detect plagiarism and contract cheating focus primarily on text-matching software, with the more recent introduction of some authorship tools. Text-matching tools are very effective for locating cut-and-paste plagiarism, where a student has lifted passages of text from other sources and added them to a document. There are instances where students attempt to obfuscate these actions, by altering things such as spacing, using white characters to make the text meaningless, and even altering character maps to confuse the text-matching software. These are more easily detected by text-matching services such as Turnitin than they once were.

Another method of detection that is gaining traction is that of authorship tools. These take some of the metadata of the document, and include factors like stylometrics and linguistics to help determine the original author of the work.

Digital Forensics tools and techniques were also reviewed as part of this research, with a view to establishing whether they could be repurposed for the detection of academic misconduct, as none of the existing work around plagiarism detection considers this area of inquiry.
There are examples of using digital forensics in computer crime, using both open-source tools like Autopsy and proprietary tools like Encase and Access Data’s FTK, and within intellectual property theft cases to a much lesser extent by examining some of the basic metadata available. However, there are no known commercial tools that allow an assessor to examine the underlying structure of a document with a view to extracting important artefacts that help to build a picture of how the document was put together, and therefore draw conclusions about the likelihood of whether the submitting student was the original author.

### 7.3 Contribution to knowledge

To further explore the digital forensics approach to academic misconduct, a software tool was developed to automatically unpack student submissions, analyse the forensic artefacts within the document that are not usually reviewed, and highlight artefacts that could indicate plagiarism or contract cheating has occurred. The tool, which has a working title of ‘Clarify’, is written in the PHP programming language, which means the results can be easily viewed as a web page. A student submission is saved to an appropriate folder on the assessor’s computer and the metadata is fully extracted using the tool. ‘Clarify’ then analyses key information from the resulting files, reporting them back to the screen for ease. There is currently no other tool that does this specifically in relation to academic integrity. Key features drawn from this analysis include:

- Extraction of rsidR edit values, including a list of unique rsidR values. These show how the document has been created or built through its edits.
- Colour visualisation of the editing that has been carried out within the document. This is very compelling evidence of whether the document is authentic, or whether it requires further investigation.
- Bar code visualisation and statistical information to show the length of text runs within the document. Large blocks of unedited text are unusual and warrant further investigation.
- Extraction of uncropped images, which can provide insight into where the image came from.
- Analysis of any hidden text, which may highlight a deliberate attempt to confuse text-matching software.
• Analysis of fonts within the document, as extensive font formatting can be indicative of copy-and-paste plagiarism.
• Anomalies in dates created and last modified.
• Anomalies in author and creator names.

This information is shared back to the assessor via a report on screen, which the option to expand the XML code and review it manually if they wish, as well as providing an overview of the findings. Included within the prototype tool are a variety of sample files to demonstrate the various outputs. There is no other tool that does this kind of analysis specifically for academic purposes that is known to the author at the time of writing, and the new techniques that ‘Clarify’ utilises are an important and useful addition to existing academic misconduct detection tools and methods. In addition, the research carried out here demonstrates a credible proof of concept that can be further developed.

The peer reviewed papers that have been published indicate that the research provides a valuable contribution to the complex field of academic integrity and academic misconduct. The student voice chapter sets the background for the research, by establishing a value base from which to carry out the detection work. Interest in the research outlining the use of digital forensics for the detection of academic misconduct, as well as in the application of cyber security techniques to academic integrity, has been consistent, and a number of people working in this area have requested to read the finished PhD thesis, as well as to test ‘Clarify’. At the time of writing, a conference presentation has been accepted for the European Conference on Ethics and Integrity in Academia 2023 to demonstrate the tool.

In addition to the above, the work highlights the importance of a whole of institution approach to academic integrity. Whilst existing text-matching and authorship software, along with the potential addition of a tool such as ‘Clarify’ to examine the intrinsic features of student submission are key, opportunities to block advertisements from essay mills, track student activity on VLE and exam portals and so on will all help to deter and detect academic misconduct.

7.4 Limitations

There are many papers on the topic of academic misconduct, contract cheating and plagiarism, and whilst a full categorisation of all articles in this field would be useful it would
be considerably time consuming. It is possible therefore that the early work of this research into student perceptions may have missed some important studies. In addition, the phrase ‘ghost writing’ was not used within the search criteria, and this may have led to some important omissions. It would be useful to take the International Center for Academic Integrity list of the top 42 Academic Integrity articles and book chapters from 1992 to 2012 (Bertram Gallant, 2012) and revisit theses with a view to systematically searching for evidence of the student voice. Combining this with further survey and article categorisation would provide a comprehensive set of data on where the student voice can be found and would ensure no major omissions have occurred. A relatively small sample of students was surveyed in the first phase of the research project (surveys). However, extensive surveys have already been done by other researchers as detailed in the literature review. In most cases these were self-selecting and yielded similar results to each other. The smaller survey group establishes that the main issues at the institution involved were not dissimilar to those found elsewhere in more extensive surveys. For the second survey, there were a smaller number of respondents, most likely due to the survey being issued online rather than in person.

In relation to the development of the detection tool, there are, of course, limitations with any tool that attempts to do this. There is always the risk that an assessor will review the scores or results and deem the student guilty of plagiarism because low numbers of edits are seen or because there are long swathes of text with limited editing. This alone is not reason for punishment, but could be the result of a student who has chosen to write chapters as individual documents, then bring them together only at the final stage of writing. Similarly, a student may choose to write something in draft, and only put it into the final document format once it has been completed. These would account for unusual artefacts (or lack of artefacts) appearing in the final document, but are not, on their own, cause for concern. Rather, these results are an excellent indication that the work needs further review, and that the student should be questioned further about the authenticity of the work, perhaps being asked to provide working drafts that preceded the final submission. The main limitation for the software development was the limited testing on authentic student work. It would have been very useful to have had a large number of genuine student submissions, written in an authentic way, to help build a baseline of what a written report or essay should look like, and how plagiarised, contracted work or work generated by AI may differ.
7.5 Recommendations and future work

‘Clarify’ is a tool in its infancy. Its aim is to demonstrate that there are many elements within a Word document that remain unexplored in terms of their ability to add evidence to the complex challenge of proving plagiarism and/or contract cheating. The tool demonstrates that this information can be extracted through an automated process and presented to the investigator in a user friendly and meaningful way, without requiring an understanding of the OOXML format nor requiring the investigator to manually review any file. However, as with any software tool in its early development, there is room for improvement. Some of the recommendations for further development are given below.

7.5.1 Genuine student work

Analysis with genuine student submissions that are not plagiarised or outsourced would help develop what an ‘authentic’ submission should look like, including things like an average number of words per rsidR run (i.e. number of words input without the need for any editing), which can then be compared with findings from the document being investigated. The challenges of gaining ethical approval for this work should not be underestimated, but it is vital future work in developing a reliable and valid tool to support academic misconduct detection.

7.5.2 Benchmarking of elements

Each of the reported values would benefit from establishing a ‘normal’ value. For example, establishing an acceptable range of text nodes compared to the number of runs would help to determine where long text runs extend beyond this range, suggesting that the work may be plagiarised (copied and pasted with no further editing). Currently ‘Clarify’ uses a barcode visual with two colours for this, but adding an ‘acceptable’ range to show only sections exceeding this would be useful. Other areas that would benefit from benchmarking include the average edit time per total word count, typical number of runs per word count, typical difference between the number of runs in the document.xml file and the settings.xml file and so on.
7.5.3 Flag system

Development of a robust and reliable flag system value would help highlight anomalies to the investigator. Alongside the improvement of the barcode system discussed above, this could include:

- Addition of the TotalTime where a disproportionately small TotalTime in comparison to the number of words in the document has been found
- Editing time of less than a day (or value chosen by the assessor)
- Revisions less than a specified number
- Appearance of rFont tags that are resetting the font back to the default (suggesting that they were originally imported in a different font). The current version of ‘Clarify’ counts all rFont tags, not just ones returning the text to the default font
- A check on the ‘created by’ name where it does not readily correlate with the submitting student. There are various reasons why this might be the case, but if it is possible to compare this to the submitting author and flag anomalies, this could provide a useful reference point for assessors.

A rudimentary flag system was tested during the investigation phase of the research (section 6.11). The first flag value was created by adding together the number of appearances of elements or attributes that were considered to be indicative of academic misconduct. These included the number of <w:sz> values and <w:shd> and could later include <w:rPr>. Next a multiplier value was given for the number of words per rsidR. Without checking a large number of genuine works, it was not possible to establish a benchmark figure for what would be acceptable, so an arbitrary figure of 25 words per rsidR was used. Submissions scoring under this value were assigned a flag score of 1, whilst submissions that showed 26 to 50 words per rsidR scored a flag score of 2, and submissions over 50 words per rsidR were given a flag score of 3. These two flag values were then combined to create a flag ‘score’.

Given the small number of works tested on, this system provided no valuable results at this point, but would be worth investigating in the future to determine whether such an approach could be beneficial.

7.5.4 Tool development areas

Other future investigations are required into the following:
• How the TotalTime is recorded if the document is left open for a long duration of time with no edits?
• Analysis of the rPr element and whether it adds any further forensic evidence.
• Automation for entire cohorts of students at the point of submission, in the same way that Turnitin is used, to flag any submissions that may require further investigation.
• Provision for other document formats using an XML structure (e.g. Google Docs, Open Office).
• Development of the tool for other Microsoft documents such as Excel and PowerPoint.
• Comparison of two documents to check for matching rsidR values that could indicate collusion.
• Review of Eric White’s Chrome extension (White, n.d.) to see if it might be possible to integrate into ‘Clarify’.

It could potentially be useful to add a ‘Tool Tip’ to the text output of the document submission. A tool tip is an on screen pop up box that appears when certain parts of the screen are hovered over with the mouse. It would be useful to use this in a variety of ways including popping up the rsidR value associated with any hovered over text. For example, where it is clear that the name of an institution has been added to a document during an editing session, but there are multiple instances of that particular editing session throughout the document, hovering over the institution name would provide the investigator with the rsidR value associated with that edit. The investigator could then toggle the corresponding rsidR value so that it highlights all text that uses the same value. If that shows that the institution name or other locally relevant variable has been amended throughout, this could indicate that the name has been changed from a previous name (perhaps another institution, or a placeholder). Another use of the tool tip feature could be to add pop up text that shows the markup belonging to that particular piece of text, which could include the rsidR value and also formatting markup, fonts, styles etc.

During the research a number of redundant formatting tags (rPr and pPr) were found to be indicative of plagiarism, where an author added information to a document with legacy formatting, or where they needed to reformat work to match the rest of the document. Large
amounts of redundant formatting tags would be unusual in an authentic document (though not impossible). Drawing these out would be useful information for the report.

Another useful addition could be ‘hiding’ all the text except for that belonging to a single edit run or rsidR value. Currently it is possible to view all the edits belonging to a single rsidR value by toggling the RSID button in the Unique rsidR values section. This highlights all text in the document belonging to that value. A development of this could be to grey out or ‘hide’ the rest of the document, making it a little easier to see what editing was done in a single session.

A further development of the tool relates to the Images section. This currently tells the investigator which images have been cropped and the amount of cropping on each side. Towards the end of the report the full image is provided along with the cropping amount. It would have been useful to extend this to shade the part of the image that was cropped out, making it very easy for the investigator to see what information had been cropped away from the original image. At the present time, this can only be done by reviewing the PDF version of the document and comparing it manually to the full image in the report. This should be possible to code with a better knowledge of PHP.

There is another XML tag that could potentially be worth adding, which is the <w:bookmarkStart w:name="_GoBack"> and <w:bookmarkEnd> tag. This element tells Word where the last edit took place, so that when the document is reopened, the cursor returns to the last edited spot to ‘Resume’ work. This information tells the investigator what the last element edited was, though how useful this is in cases of academic misconduct is questionable and therefore this was not added to the program.

Finally, it could be possible to develop student ‘profiling’. Belvisi et al. (2020) point to this in their research, and whilst in their case it relates to stylistic features, it may be possible to create a metadata profile of a student, such that deviations from this profile could suggest that further investigation is needed.

### 7.6 Final thoughts

Digital forensics is a field that is currently underexplored in relation to academic integrity. There are a handful of very good papers on this matter, but very few focusing on these techniques in relation to academic misconduct. Given that plagiarism and contract cheating
challenge educators’ moral and ethical standing, the addition of automated tools that can help assessors with the ever-increasing challenges to academic integrity could be useful. ‘Clarify’ has gone some way into establishing which artefacts may be useful for assessors to be aware of, and with more testing, it may be possible to establish benchmarks against which documents can be checked. The aims of the research were to find out:

1. What are student perspectives towards academic integrity? and
2. Can plagiarism and contract cheating detection tools be improved using digital forensics techniques?

In respect of point 1, the literature review and primary research around the student voice provided good insights into student perspectives, and this resulted in publication of a peer reviewed book chapter as well as an international conference presentation and internal staff workshops. This area of research related to prevention and penalty which were considered through the lens of the student and their perceptions of what academic integrity and academic misconduct are. In an ideal world, academic institutions would put all their energy into this phase, and the need for any other stages would be negated. However, this is not realistic, and the research indicated that students’ understanding of academic integrity is limited. Thus, the focus of the research turned to point 2, that of detection, with development of a novel approach to detecting plagiarism and contract cheating through a software tool with the working title ‘Clarify’. The tool demonstrates that new detection techniques are possible, by going beyond simple text focused analysis of document to explore the metadata of a document. Presentations on the digital forensics techniques used by ‘Clarify’ have been delivered at several international conferences as well as internally at the researcher’s institution. In addition, two journal articles and two book chapters have been published, with a third book chapter accepted for print and the work to date clearly establishes the value of such a tool, and that further development and integration into existing authorship / text-matching tools could add valuable evidence for an academic misconduct panel.
References


AtomikRadio, (2015). ‘I can't figure out how this student beat Turnitin. Any insight?’, Reddit, 28 April. Available at: https://www.reddit.com/r/college/comments/3461m0/i_cant_figure_out_how_this_student_beat_turnitin/ (Accessed 28/06/21)


Understanding academic integrity and plagiarism in the digital age  Clare Johnson


Fishman., T., (2009). ‘We know it when we see it’ is not good enough: Toward a standard definition of plagiarism that transcends theft, fraud, and copyright. In Proceedings 4th Asia Pacific Conference on Educational Integrity (4APCEI’09). 5.

Understanding academic integrity and plagiarism in the digital age

Clare Johnson


Interpol, (2022). Digital Forensics, Available at: https://www.interpol.int/How-we-work/Innovation/Digital-forensics


Johnson, C., & Davies, R., (2020b). Plagiarism from a digital forensics perspective, In Integrity in Education for Future Happiness (pp. 78-89), Brno: Mendel University in Brno


Johnson, C., Reddy, M., & Davies, R., (2023 in print), Cyber security: the next defence for academic misconduct, in Handbook of Academic Integrity, 2nd Edition


Understanding academic integrity and plagiarism in the digital age  
Clare Johnson


Measure Learning, (2022). Advanced exam security, Available at: https://www.measurelearning.com/services/advanced-exam-security


Quality Assurance Agency for Higher Education, (2022b). ‘Subject Benchmark Statements’, Available at: https://www.qaa.ac.uk/quality-code/subject-benchmark-statements


University of South Wales, (2022). Academic Integrity Statement for Students, Available at: [https://registry.southwales.ac.uk/student-regulations/academic-misconduct/](https://registry.southwales.ac.uk/student-regulations/academic-misconduct/)


Understanding academic integrity and plagiarism in the digital age Clare Johnson

References for sample essays

Plagiarism Essay


Python Essay


Raspberry Pi Essay

GitHub, (2020). Pi-hole / setup.py, Available at: https://github.com/pi-hole/pi-hole/blob/master/setup.py, accessed 20/02/20

Human Computer Interface Essay


IoT Essay


Red Team Blue Team Essay


**Social Engineering Essay**


**Blockchain Essay**


**Online Discussion Fora Essay**


**Appendix A: Plagiarism tools (intrinsic and extrinsic)**

<table>
<thead>
<tr>
<th>Tool Name</th>
<th>Text-matching using sources such as journal databases, web searches etc (Extrinsic)</th>
<th>Authorship tools such as stylometrics (Intrinsic)</th>
<th>Forensic analysis using xml</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chowdhury &amp; Bhattacharyya’s list of tools</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SafeAssignment</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Docol</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urkund</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Copycatch</td>
<td>E</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>WCopyfind</td>
<td>E</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Eve2</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPSP</td>
<td>I</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MOSS</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JPlag</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CopyScape</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DOC Cop</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ephorus</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lhenticate</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarism Detect</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Exactus Like</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DupliChecker</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarisma</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarism Checker</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagium</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PlagTracker</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Quetext</td>
<td>E</td>
<td>I</td>
<td></td>
</tr>
<tr>
<td>Turnitin</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Viper</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Maulik</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarism Scanner</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hawk Eye</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Code Match</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SID-software Integrity Diagnosis System</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SIM</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YAP3</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PlagScan</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Foltýnek et al’s list of tools (2020) – only additional tools to the above list are included</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Akademia</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPV</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PlagAware</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plagiarism Software</td>
<td>Couldn’t locate</td>
<td></td>
<td></td>
</tr>
<tr>
<td>StrikePlagiarism</td>
<td>E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unicheck</td>
<td>E</td>
<td>I</td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Greyscale versions of images

This appendix provides greyscale / patterned versions of image outputs to aid visualisation in printed copies of the text. The original colour versions are inline. Patterns have been manually applied in Photoshop replacing matching colours (e.g. a dot pattern may replace all appearances of the identical shade of red within the image).

Figure 2. 1: Visualisation of the original version of the Author Work Example Extract, written by the first party, Johnson & Davies 2020a (Greyscale)

Figure 2. 2: Visualisation of the copied and edited work of the Author Work Example Extract, as viewed from the second party, Johnson & Davies 2020a (Greyscale)
Appendix B: Greyscale versions of images
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Figure 2.3: Visualisation of the Contracted Author's work (colour), Johnson & Davies 2020a (Greyscale)

Figure 2.4: Visualisation of the student submission (colour), Johnson & Davies 2020a
Appendix B: Greyscale versions of images
Understanding academic integrity and plagiarism in the digital age  Clare Johnson

Figure 2. 5: Colour coded visualisation of text showing a normal editing pattern (greyscale)

Figure 2. 6: Colour coded visualisation of text demonstrating an unusual editing pattern (greyscale)

Figure 2. 7: Colour coded visualisation of text from sample L4002 (greyscale)
Appendix B: Greyscale versions of images
Understanding academic integrity and plagiarism in the digital age

Figure 2. 8: Visual representation of L5002 suggestive of patchwriting (greyscale)

Figure 2. 9: Visual representation of L4005, in class test (greyscale)

Figure 2. 10: Visual representation of L4004, where a more consistent sequence of editing can be seen (greyscale)

Figure 2. 11: Visual representation of extract of L7004, showing large blocks of unedited text (greyscale)
Appendix C: Investigations

Investigation 1: Creating a simple Word document

- A new document was created in Word, using File, New Document
- A title “Test 1” was added, and a new line created by pressing the Enter key
- The document was saved with the file name “test1.docx”
- The text “Content 1” was added, and a new line created (Enter)
- The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”, and a new line created (Enter)
- The document was saved again.
- The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed.

This yielded the following results:

![Figure 3.1: Visual of how rsidR values are built in Investigation 1](image)

The red bar | indicates a save point.

In table form, this is the process that was followed:
Table 3.1: Observations from investigation 1

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A new document was created in Word, using File, New Document</td>
<td></td>
</tr>
<tr>
<td>• A title “Test 1” was added</td>
<td>rsidRDefault: 00C90590 was generated</td>
</tr>
<tr>
<td>• The document was saved with the file name “test1.docx”</td>
<td>rsidR: 00026756 was generated</td>
</tr>
<tr>
<td>• The text “Content 1” was added</td>
<td>rsidRDefault: 00143CFB was generated and rsidR adopted the value of the previous rsidRDefault element</td>
</tr>
<tr>
<td>• The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the word “inline”</td>
<td>rsidRDefault remained as 00143CFB and rsidR adopted the value of the previous rsidRDefault element (which then matched the current rsidRDefault value) No rsidRPr was generated as formatting was carried out inline</td>
</tr>
<tr>
<td>• The document was saved again.</td>
<td></td>
</tr>
<tr>
<td>• The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed.</td>
<td>New rsidRDefault: 00580F26 was generated (new save point) and rsidR adopted previous rsidRDefault value. rsidRPr took current value of rsidR element.</td>
</tr>
<tr>
<td></td>
<td>New rsidRPr value of 00580F26 (which comes from the previous rsidRDefault value) was added to the newly formatted words to show that the content had been edited / changed retrospectively – this suggested that the original text took the rPr value to match the rsidR value for the run, and that the subsequent edit needed a new rPr value to show that it was edited separately, and therefore took the rsidRDefault value for the run (so it was different to the existing rPr value).</td>
</tr>
<tr>
<td>• FOOTER SECTION</td>
<td>The header/footers section (w:sectPr) contained values for rsidR of 00143CFB (rsidR value from final &lt;w:p&gt; element) and rsidRPr of 00143CFB</td>
</tr>
</tbody>
</table>

rsidRDefault: 00C90590 was generated
rsidR: 00026756 was generated
rsidRDefault: 00143CFB was generated and rsidR adopted the value of the previous rsidRDefault element
rsidRDefault remained as 00143CFB and rsidR adopted the value of the previous rsidRDefault element (which then matched the current rsidRDefault value) No rsidRPr was generated as formatting was carried out inline
New rsidRDefault: 00580F26 was generated (new save point) and rsidR adopted previous rsidRDefault value. rsidRPr took current value of rsidR element. New rsidRPr value of 00580F26 (which comes from the previous rsidRDefault value) was added to the newly formatted words to show that the content had been edited / changed retrospectively – this suggested that the original text took the rPr value to match the rsidR value for the run, and that the subsequent edit needed a new rPr value to show that it was edited separately, and therefore took the rsidRDefault value for the run (so it was different to the existing rPr value).
Appendix C: Investigations
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Investigation 2: Adding formatting to a document

- A new document was created in Word, using File, New Document
- The file was saved with the name “test2.docx”
- A title “Test 2” was added, and a new line created (Enter)
- The document was saved again
- The text “Content 1” was added, and a new line created (Enter)
- The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”, and a new line created (Enter)
- The document was saved again.
- The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was saved and closed.

This yielded the following results:

![Figure 3.2: Visual of how rsidR values are built in Investigation 2](image-url)

Figure 3.2: Visual of how rsidR values are built in Investigation 2
To clarify this:

Table 3.2: Observations from investigation 2

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A new document was created in Word, using File, New Document</td>
<td></td>
</tr>
<tr>
<td>• The document was <strong>saved</strong> with the file name “test2.docx”</td>
<td></td>
</tr>
<tr>
<td>• A title “Test 2” was added</td>
<td></td>
</tr>
<tr>
<td>• The document was saved again</td>
<td></td>
</tr>
<tr>
<td>• The text “Content 1” was added</td>
<td></td>
</tr>
<tr>
<td>• The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”</td>
<td></td>
</tr>
<tr>
<td>• The document was <strong>saved</strong> again.</td>
<td></td>
</tr>
<tr>
<td>• The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. The document was <strong>saved</strong> and closed.</td>
<td></td>
</tr>
<tr>
<td>• <strong>FOOTER SECTION</strong></td>
<td></td>
</tr>
<tr>
<td>• The header/footer section (w:sectPr) contained values for rsidR of 004F7894 (rsidR value from final &lt;w:p&gt; element) and rsidRPr of 004F7894</td>
<td></td>
</tr>
</tbody>
</table>

rsidRDefault changed to 004F7894 and rsidR adopted the value of the previous rsidRDefault element

rsidRDefault remained as 004F7894 and rsidR adopted the value of the previous rsidRDefault element (which then matched the current rsidRDefault value) No rsidRPr required as formatting carried out inline

rsidRPr took current value of rsidR element.

New rsidRDefault value of 00B635EF was recorded (new save point) and rsidR adopted previous rsidRDefault value. rsidRPr took current value of rsidR element.

New rsidRPr value of 00B635EF (which comes from the previous rsidRDefault value) was added to the newly formatted words to show that the content had been edited / changed retrospectively
Investigation 3: Editing a document

Investigation 4: Copy, Paste and Edit

- The file was saved with the name “test3.docx”
- A title “Test 3” was added, and a new line created (Enter)
- The document was saved again
- The text “Content 1” was added, and a new line created (Enter)
- The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline”, and a new line created (Enter)
- The document was saved again.
- The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold.
- The title was edited to add the words “- Editing the title” after “Test 3”
- The document was saved and closed.

This yielded the following results:

Figure 3.3: Visual of how rsidR values are built in Investigation 3
To clarify this:

### Table 3.3: Observations from Investigation 3

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
</table>
| • A new document was created in Word, using File, New Document  
• The document was saved with the file name “test3.docx” |  
rsidRDefault: 00DA7DC6 was generated and rsidR: 00026756 was generated |
| • A title “Test 3” was added  
• The document was saved again | rsidRDefault changed to 00F23316 and rsidR adopted the value of the previous rsidRDefault element (00DA7DC6) |
| • The text “Content 1” was added | rsidRDefault remained as 00F23316 and rsidR adopted the value of the previous rsidRDefault element (which then matched the current rsidRDefault value) No rsidRPr required as formatting carried out inline |
| • The text “Content 2 – bold applied inline” was added, with the bold being applied by turning on bold formatting before the word “bold” and turned off after the work “inline” |  
New rsidRDefault value of 005C4453 recorded (new save point) and rsidR adopted previous rsidRDefault value (00F23316). rsidRPr took the current value of rsidR element.  
New rsidRPr value of 005C4453 (which came from the previous rsidRDefault value) added to the newly formatted words to show that the content had been edited / changed retrospectively |
| • The document was saved again. |  
A new rsidR value of 005C4453 was added around the inserted text which inherited the rsidRDefault value from the paragraph which was being editing previously (i.e. the final paragraph of the document) |
| • The text “Content 3 – bold applied retrospectively” was added, then the words “bold applied retrospectively” highlighted and made bold. | The header/footer section (w:sectPr) contained values for rsidR of 00F23316 (rsidR value from final <w:p> element) and rsidRPr of 00F23316 |
| • The title was edited to add the phrase “- Editing the title” after “Test 3”  
• The document was saved and closed |  
A new rsidR value of 005C4453 was added around the inserted text which inherited the rsidRDefault value from the paragraph which was being editing previously (i.e. the final paragraph of the document) |

FOOTER SECTION

The header/footer section (w:sectPr) contained values for rsidR of 00F23316 (rsidR value from final <w:p> element) and rsidRPr of 00F23316
Investigation 4: Copy, Paste and Edit

- A new document was created and saved with the file name “test4.docx”
- Content from the test3 file was copied and pasted into test4 (note the title was left as Test 3)
- The file was saved
- The title was edited to delete the words ‘Editing the file’
- The file was saved and closed

Figure 3.4: Visual of how rsidR values are built in Investigation 4

To clarify this:

Table 3.4: Observations from Investigation 4

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A new document was created in Word, using File, New Document</td>
<td>rsidRDefault: 009B0641; rsidR: 009B0641; and a new element of rsidP with the value 009B0641 was generated</td>
</tr>
<tr>
<td>• The document was saved with the file name “test4.docx”</td>
<td></td>
</tr>
<tr>
<td>• The entire contents of test3.docx were copied and pasted into test4.docx and the file saved</td>
<td></td>
</tr>
<tr>
<td>• The text “- Editing the title” was deleted and the document saved and closed</td>
<td>All rsidR values throughout the document became 009B0641 with the exception of the final paragraph element containing no content which had the rsidR value of 00026756 and a new rsidRDefault value of 004F5EA3</td>
</tr>
<tr>
<td>• FOOTER SECTION</td>
<td>The header/footer section (w:sectPr) contained values for rsidR of 00026756 (which appeared to come from the first rsidR value of the original document. There was no rsidRPr value</td>
</tr>
</tbody>
</table>
Investigation 5: Exploring the sequence of document edits

- The file was saved with the name “test5.docx”
- A title “Test 5” was added, a new line (Enter) created and the file saved
- The text “Content 1” was added, a new line (Enter) created and the file saved
- The text “Content 2” was added, a new line (Enter) created and the file saved
- The text “Content 3” was added and the file saved
- The title was edited to add the words “– Editing the title” after “Test 5” and the file saved
- The phrase “– this phrase was added retrospectively” was added after “Content 2” and the file saved
- The phrase “This phrase was inserted retrospectively -” was inserted before “Content 3” (note the document was not saved here)
- Finally, a new heading was added on a new line at the top of the document with the text “Plagiarism Experiments”
- The document was saved and closed.

This yielded the following results:

NOTE: the device used for creating these experiments was replaced prior to Investigation 5 (filename: test5.docx):

Figure 3.5: Investigation 5 diagram
### Table 3.5: Observations from Investigation 5

<table>
<thead>
<tr>
<th>ACTION</th>
<th>Observations</th>
</tr>
</thead>
</table>
| • A new document was created in Word, using File, New Document  
  • The document was saved with the file name “test5.docx” | rsidRDefault: 00AA7542 was generated and rsidR: 00AA7542 was generated |
| • A title “Test 5” was added  
  • The document was saved again | rsidRDefault: 00414A12 was generated and rsidR: 00FF7eD2 was generated |
| • The text “Content 1” was added and the document saved | rsidRDefault changed to 00032912 and rsidR adopted the value of the previous rsidRDefault element (00414A12) |
| • The text “Content 2” was added and the document saved | rsidRDefault changed to 007A49C7 and rsidR adopted the value of the previous rsidRDefault element (00032912) |
| • The text “Content 3” was added and the document saved. | rsidRDefault changed to 00AA7542 recorded and rsidR changed to 00510958 – note that the inserted phrase took the rsidRDefault value of the preceding paragraph, not the original text of ‘Content 3’ |
| • The title was edited to add the phrase “- Editing the title” after “Test 5”  
  • The document was saved | A new rsidR value of 006B2A4D was added around the inserted text – in this case this was a newly generated rsidR and was not inherited from the previous paragraph being edited (unlike in the previous example – perhaps because the save sequence was different) |
| • The phrase “ – this phrase was added retrospectively” was added after “Content 2” and the file saved | A new rsidR value of 0012522A was added around the inserted text – again this was NOT inherited from elsewhere |
| • The phrase “This phrase was inserted retrospectively - ” was inserted before “Content 3” (note the document was not saved here) | A new rsidR value of 007A49C7 was added around the inserted text, this was inherited from the preceding paragraph’s rsidRDefault value |
| • Finally, a new heading was added on a new line at the top of the document with the text “Plagiarism Experiments”  
  • Document was saved and closed | A new rsidR value of 00AA7542 was added around the inserted text, and the paragraph enclosing also had the rsidRDefault value of 00AA7542 (which matched the rsidRDefault value of the paragraph in which the preceding edit took place) |
| • FOOTER SECTION | The header/footer section (w:sectPr) contained values for rsidR of 007A49C7 (rsidR value from the paragraph which contains the text that was edited last of all) |
Investigation 7: Multiple edits in a single editing session

The document test6.docx was reopened and saved as test7.docx. Each appearance of the word ‘Content’ was replaced with the word ‘Heading’ in one editing session. The text ‘Test 5’ was edited to ‘Test 7’. At the end of the document, a final sentence was added with the text ‘End of document’. The document was saved.

This yielded the following results:

1. The paragraph containing the text “Test ” retained the same rsidR and rsidRDefault values as in test6. Where the test number was changed to a “7” it was wrapped in an edit run (<w:r>) with the rsidR value 00C411A9 (new value).

2. Where “Content 1” was changed to “Heading 1”, the paragraph element (<w:p>) containing the text retained the rsidR value from the previous test of 00414A12, but the rsidRDefault value changed from 00032912 in test6 to 00C411A9 in test7 (which matched the rsidR value of the editing run in part 1). The text “Heading” therefore picked up the paragraph rsidR and rsidRDefault values, but the original number “1” retained the rsidR value of 00032912 (which was the rsidRDefault paragraph value from test6 where this text was originally created).

3. Where “Content 2” was changed to “Heading 2”, the paragraph element containing the text retained the rsidR value from the previous test of 00032912, but the rsidRDefault value changes from 007A49C7 in test6 to 00C411A9 in test7 (which matched the rsidR value of the editing run in parts 1 and 2). The text “Heading” therefore picked up the paragraph rsidR and rsidRDefault values, but the original number “2” retained the rsidR value of 007A49C7 (which was the rsidRDefault paragraph value from test6 where the original text was created).

4. Where “Content 3” was changed to “Heading 3”, the paragraph element containing the text retained the rsidR value of 007A49C7 and the rsidRDefault value of 00AA7542 from the previous test. The text “Heading” was wrapped in a separate edit run with an rsidR value of 00C411A9 which matched the rsidRDefault value from the edits carried out so far, and the “3” had an rsidR value of 00510958 (which was the same as the rsidR value for the original run in test6. This was a slightly different
mark-up to the previous edits, most likely because the paragraph had already been edited in test6.

5. The final (added) paragraph “End of document” had matching rsidR and rsidRDefault values of 00C411A9 – this was the new edit run value seen from part 1, 2, 3 and 4 (reflecting that all the edits were carried out in the same run.
### Table 3.6: Observations from investigation 7

<table>
<thead>
<tr>
<th>Edit Number</th>
<th>Action</th>
<th>rsidR Value</th>
<th>rsidRDefault Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Empty doc saved</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
| 2           | Test 5 added  
New values generated | FF7ED2      | 414A12            |
| 3           | Content 1 added  
rsidR inherited rsidRDefault value from previous paragraph | 414A12      | 032912             |
| 4           | Content 2 added  
rsidR inherited rsidRDefault value from previous paragraph | 032912      | 7A49C7             |
| 5           | Content 3 added  
Generated a new rsidR value because the paragraph was later edited | 510958      | AA7542             |
| 6           | Editing the title added  
Generated a new rsidR value because this was a separate edit | 6B2A4D      | 414A12             |
| 7           | This phrase(2)... added  
Generated a new rsidR value | 12522A      | 7A49C7             |
| 8           | This phrase(3)... added  
rsidR inherited rsidRDefault value from previous paragraph | 7A49C7      | AA7542             |
|             | Plagiarism Experiments added  
rsidR inherited rsidRDefault value from last paragraph edited | AA7542      | AA7542             |
| FOOTER      | Appeared to pick up rsidR value from final paragraph in document | 7A49C7      | 7A49C7             |
Investigation 9: Effect of no user invoked save on a document

The purpose of this investigation was to explore the rsidR and rsidRDefault values generated when various methods of adding text to the document were carried out but no save was invoked. These files used were saved as testa1 and testa2:

- A blank document was created and saved. The rsidR and rsidRDefault values were noted
- Content was added used =rand() and then using other methods such as typing directly into the document, copy-and-paste, but all within a single save

Results:

- rsidR and rsidRDefault values in testa1 were 00A07304 (same as above).

After =rand() content added (testa2):

- Every paragraph had the rsidRDefault value of 00813A78 including the last paragraph (which had NO content)
- Every run of text had the rsidR value of 00813A78 inherited from the paragraph Default value EXCEPT the final paragraph containing text, which had an rsidR value of 00A07304). This value matched the rsidR value from the previous saved version of the document.

Investigation 10: Entering text manually in a single session

Text was typed manually (3 paragraphs) into a blank document with the filename testb:

- Initial rsidR value was 00A07304 (this echoed previous experiments where the first rsidR value was the same for all documents created on the same device).
- The rsidRDefault value was 00AA424B for all three paragraphs (all created in a single editing session).
- The second and third paragraphs had an rsdiR value of 00AA424B, which was the same as the rsidRDefault value for all three paragraphs
- The footer section also had an rsidR value inherited from the rsidRDefault value throughout of 00AA424B.
Investigation 11: Entering text manually in a single session (repeat)

Text was typed manually (3 paragraphs) into a blank document with the filename testc2:

- Initial rsidR value was 00A07304 (this echoes previous experiments where the first rsidR value was the same for all documents created on the same device).
- The rsidRDefault value was 00AF4E21 for all three paragraphs (all created in a single editing session).
- The second and third paragraphs had an rsdiR value of 00AF4E21, which was the same as the rsidRDefault value for all three paragraphs.
- The footer section also had an rsidR value inherited from the rsidRDefault value of 00AF4E21.

Investigation 12 – Editing an existing document

The file testc2 was reopened and five edits were carried out (3 formatting changes, and two text changes including adding words and changing words) all in one session, then saved as test3c.

- Initial rsdiR value and rsidRDefault values were 00444C92 (changed from previous examples) – This was a newly inserted paragraph.
- The rsidRDefault value for all other paragraphs was 00AF4E21 as in previous test so remained unaltered.
- The second and third paragraphs took the rsidR value of 00AF4E21 (unchanged).
- The edited text (word change) was enclosed in a new rsidR run value of 00444C92 (this run value was inherited from the rsidRDefault value).
- The format edits generated a new rsidRPr value (previously there were no rPr values) – all three were the same and adopted the rsidRDefault value of 00444C92.
- The footer section had an rsidR value inherited from the original rsidRDefault value of 00AF4E21 (remaining unchanged from the original).
Investigation 13 – Adding text to a blank document, then editing a single word

Random text was added to a document (5 paragraphs) and the document saved as testd.

- All rsidR and rsidRD values are identical

When one word was altered (the word ‘change’ becomes ‘alter’):

- The w14:textId value changed from 77777777 to 687E170C and the word ‘alter’ was wrapped in a separate run with the rsidR value of 00595ED7.
- No other changes

Reviewing the settings.xml for this document, 5 unique rsidR values exist. These were:

- 00595ED7
- 00693249
- 00A07304 – this was the device identifier
- 00BC5912
- 00E70347
Investigation 14 – Mapping the rsidR values (TestF)

Tests were carried out (TestF1, TestF2, TestF3 and TestF4) to map the evolution of rsidR values in a longer document. A combination of manually typed additions to the file, as well as insertions of random text using =rand() and copied and pasted text were used as detailed below:

- A new blank document was created. The text ‘Student Name’ was added to the start of the document. Random text was entered using =rand(). The document was saved with the file name TestF1
- The document was duplicated (to facilitate analysis) and the second version renamed to TestF2
- TestF2 was opened and three edits made (two additions of text: “Images can be inserted where required” and “Design ideas can make your presentation really professional” and one formatting change – bold of ‘Student Name’). The document was saved
- The document was duplicated (to facilitate analysis) and the third version renamed to TestF3
- TestF3 was opened and a passage of text copied (“Microsoft Word is a word processor…. for Xenix systems”) from the Internet was added. The file was saved.
- The document was duplicated (to facilitate analysis) and the third version renamed to TestF4
- TestF4 was opened and a passage of text copied from the Internet was added (“Collaborate with others…”). This was re-formatted using the font face and font size tools. The file was saved

This yielded the following results:
Appendix C: Investigations
Understanding academic integrity and plagiarism in the digital age

Figure 3.6: Observations from Investigation 14

*Clear Line Spaces not deliberately added
The following observations were made from this experiment:

- All edits carried out in edit session 1 took the initial rsidRDefault of 002C66F6;
- Text added manually in edit session 2 was given new and unique rsidR and rsidRDefault values;
- The formatting change in edit session 2 (bold) was allocated the rsidRDefault value from edit session 2 despite the text belonging to edit session 1;
- Where text was pasted from the Internet in edit session 3, new and unique rsidR, rsidRDefault and rsidRPr values were added. The inclusion of an rsidRPr value appears to be because this is in a different font to the rest of the document (hence requires formatting information to be added);
- The text pasted in at edit session 4 was also given new and unique rsidR, rsidRDefault and rsidRPr values.
### Appendix D: Table of forensic artefacts found in student submissions

#### Table 4. Forensic artefacts found in student submissions

<table>
<thead>
<tr>
<th>Reference</th>
<th>Turnitin Score</th>
<th>Word count</th>
<th>Revisions</th>
<th>Total time (time open for editing in mins)</th>
<th>Average words per run</th>
<th>Number of paragraphs in document.xml</th>
<th>Ratio of words to rsidR values</th>
<th>Unique rsidR values in settings.xml</th>
<th>Unique rsidR values in document.xml</th>
<th>w:Font ascii</th>
<th>w:rFont</th>
<th>w:shd clear</th>
<th>w:sz</th>
</tr>
</thead>
<tbody>
<tr>
<td>L4002</td>
<td>49%</td>
<td>2186</td>
<td>256</td>
<td>2178</td>
<td>9.18</td>
<td>70</td>
<td>11.63</td>
<td>79</td>
<td>138</td>
<td>0</td>
<td>33</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>L4003</td>
<td>45%</td>
<td>2524</td>
<td>26</td>
<td>6399</td>
<td>6.41</td>
<td>146</td>
<td>8.09</td>
<td>68</td>
<td>93</td>
<td>2</td>
<td>77</td>
<td>117</td>
<td>31</td>
</tr>
<tr>
<td>L4004</td>
<td>42%</td>
<td>2176</td>
<td>14</td>
<td>200</td>
<td>9.46</td>
<td>119</td>
<td>10.07</td>
<td>50</td>
<td>70</td>
<td>570</td>
<td>11</td>
<td>348</td>
<td>4</td>
</tr>
<tr>
<td>L4005</td>
<td>79%</td>
<td>417</td>
<td>2</td>
<td>0</td>
<td>3.16</td>
<td>80</td>
<td>3.28</td>
<td>29</td>
<td>37</td>
<td>69</td>
<td>60</td>
<td>202</td>
<td>17</td>
</tr>
<tr>
<td>L4008</td>
<td>53%</td>
<td>3411</td>
<td>1</td>
<td>139</td>
<td>8.86</td>
<td>81</td>
<td>15.97</td>
<td>22</td>
<td>22</td>
<td>842</td>
<td>31</td>
<td>461</td>
<td>0</td>
</tr>
<tr>
<td>L5001</td>
<td>50%</td>
<td>2970</td>
<td>10</td>
<td>11576</td>
<td>10.17</td>
<td>55</td>
<td>12.53</td>
<td>30</td>
<td>33</td>
<td>627</td>
<td>37</td>
<td>341</td>
<td>0</td>
</tr>
<tr>
<td>L5002</td>
<td>46%</td>
<td>1558</td>
<td>5</td>
<td>108</td>
<td>12.08</td>
<td>27</td>
<td>26.41</td>
<td>12</td>
<td>14</td>
<td>13</td>
<td>12</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>L7001</td>
<td>51%</td>
<td>1190</td>
<td>4</td>
<td>155</td>
<td>5.41</td>
<td>38</td>
<td>7.26</td>
<td>11</td>
<td>11</td>
<td>342</td>
<td>4</td>
<td>184</td>
<td>0</td>
</tr>
<tr>
<td>L7003</td>
<td>46%</td>
<td>3600</td>
<td>60</td>
<td>3132</td>
<td>9.57</td>
<td>128</td>
<td>13.90</td>
<td>69</td>
<td>101</td>
<td>90</td>
<td>33</td>
<td>194</td>
<td>0</td>
</tr>
<tr>
<td>L7004</td>
<td>74%</td>
<td>18706</td>
<td>27</td>
<td>88</td>
<td>12.27</td>
<td>741</td>
<td>24.61</td>
<td>12</td>
<td>33</td>
<td>2022</td>
<td>0</td>
<td>1260</td>
<td>607</td>
</tr>
</tbody>
</table>
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

This appendix provides the initial contract cheating case that sparked the creation of ‘Clarify’. The files have been redacted to ensure anonymity, but the ‘Clarify’ reports are based on the original files submitted. Included in this appendix are:

- The Contracted Author’s work, analysed by ‘Clarify’ with associated outputs
- The essay submitted by the student, analysed by ‘Clarify’ with associated outputs

In the Contracted Author document, many colours can be seen indicating lots of editing has taken place throughout the document’s creation. Small edits can also be seen throughout and this shows a fairly normal pattern of editing. Contrast this with the student submission, which shows large blocks of colour with little editing. It was observing this difference in the digital forensics exploration of the file that triggered initial investigations into what the forensic artefacts meant and this technique was later refined and developed to create ‘Clarify’.

Despite the contracted author’s work showing lots of editing throughout, there are a substantial amount of font changes (currently shown in full on the output), which is suggestive of extensive reformatting, perhaps indicating that the work was not original even when provided by the contracted author.

Note that names have been obfuscated for anonymity reasons.
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Contracted Author’s work output from ‘Clarify’

---

### Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>INTRODUCTION</td>
<td>2</td>
</tr>
<tr>
<td>TECHNOLOGY, THE INTERNET AND THE WEB</td>
<td>2</td>
</tr>
<tr>
<td>CYBER ATTACK</td>
<td>3</td>
</tr>
<tr>
<td>OF CRIME AND CRIMINALS</td>
<td>4</td>
</tr>
<tr>
<td>CYBER SECURITY</td>
<td>5</td>
</tr>
<tr>
<td>DIGITAL FORENSICS AND EVIDENCE COLLECTION</td>
<td>5</td>
</tr>
<tr>
<td>CONFIDENTIALITY INTEGRITY AND AVAILABILITY</td>
<td>7</td>
</tr>
<tr>
<td>CONCLUSION</td>
<td>8</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>9</td>
</tr>
</tbody>
</table>

---

## Results Section

### Document.xml Output

Show or hide the document.xml code here

The following section displays all text elements within the document. Blocks belonging to one text run are separated by a red pipe (|). Edit runs (rsidR) can be seen by using the toggle switches in the rsidR values section.

```xml
ACADEMIC PAPER | CURRENT CHALLENGES IN DIGITAL FORENSIC INVESTIGATION | By
THE INTERNET AND THE WEB | CYBER ATTACK | OF CRIME AND CRIMINALS
D | CYBER SECURITY | DIGITAL FORENSICS AND EVIDENCE COLLECTION
CONFIDENTIALITY INTEGRITY AND AVAILABILITY | CONCLUSION | REFERENCES
We live in a dynamic world. Things do not stay the same for a long time. In the last few decades, technology has brought an advancement I never experienced before and most recently, this change has intensely transformed the way we live our lives. All around the world, governments are gearing up to tackle the next threat to global security—cyber attacks. I launch programmes to support key infrastructure sectors.
```

229
the European cyber centre I support member states to respond to cyber crimes | Combating cyber crime will require a conclusive | formalised | international network | whereby law enforcement agencies will go beyond cyber space to prosecute those engaged in illegal cyber crimes. Criminal techniques are constantly changing but their goals remain the same (destroy or use hacking to protect a status quo). At the moment, I organisations | I seem to be mostly affected by cyber attacks, partly because the hackers are doing it majorly for the monetary gain and partly because there are so many vulnerable | organisations | I lost there. The supply chain system where people and | organisations | I are involved in the activities, sharing information and resources to move product from customers to suppliers using technology | I left major businesses vulnerable to attacks. The mechanisms of supply chain enabled information and security to be shared across | organisations | I making a chain of | organisations | I to be potentially as strong as the weakest member in the chain. To make it more clearer, in a setting where an organisation outsources their process, IT management or security management to a Managed Service Provider or a 3rd | Party, the security system of the client organisation | I is only as strong as the security system and processes put in place by the 3rd | Party. This is because all the client | organisations | infrastructure, system, data are vulnerable or accessible to the 3rd | Party and it only takes one vindictive employer, an intelligent hacker to attack the 3rd | Party, all the client | organisations | subscribed to the services of the 3rd | Party would come crumbling down like a deck of cards. | TECHNOLOGY, THE INTERNET AND THE WEB | Technology, the internet and the web form the infrastructure of the Internet of Things (IOT) | I have we today. The Internet evolves as the hardware, the | things | I, while the Internet can be seen as the vehicle. The web is the channel through which the Internet | I can and which | enable | Internet | communication and networking. | These three together have transformed the different facets of human existence: travel, business, communication and relationship. The Internet of Things has pervaded all | I start of human endeavour; the economic, social, technical etc. | Everyday living and activities have been pervaded by the Internet of Things. The health sector, the banks, the schools, architecture, legal practice, the government and other sectors. All capabilities are affected. Basically all things operating without the Internet before are now | I being connected to the internet for better living and advancement. Daily inaction is getting the ‘checkers’ moment where they are they discovering how better to create new things or improve old things with the internet. Information technology is everywhere, it is not going to go away I, it can only evolve in its use and its applications. Elasticity | I, water, transportation, the media, logistics etc are all depending on the use of ICT. | As much and as wide as the Internet of Things has proliferated, so has the activities of hackers been intensified and become more sophisticated. News are broadcast everyday of how the hackers are ravaging the society, disrupting performance and productivity. Various researches, policies and acts have been declared as war against the dark world. Every day, enterprises are investing millions of dollars to combat this heinous monster. It has been predicted that worldwide spending will climb to 90 billion in 2018. | Rose, S. (Ed), Eldridge, L., Chapman, S. (2015). p.11. | I Schick, 2017, p.11. | I Jones, David (2016, p.13) | Cyber attack Cybercrime is any illegal crime or transaction done over the Internet or in the cyberspace and which is targeted at an individual, a business, government or the public. | I can come from a variety of sources, including: foreign nations engaged in espionage and information warfare; criminals; hackers; virus writers; and disgruntled employees and contractors working within an organisation. There are two categories of cyber attack: I Unintentional | I Unintentional threats include both targeted and non-targeted attacks. | I Unintentional I Unintentional threats can be caused by inadvertent or untrained employees, software upgrades, maintenance procedures and equipment failures that inadvertently disrupt computer systems or corrupt data. | Targeted | I Non-targeted I Cyber attack I can take two forms, I A targeted attack occurs when the intended target of the attack is uncertain, such as when a virus, worm, or malware is released on the Internet with no specific target. I Hackers come in every shapes and sizes | I some to highlight the weakness of computing system, some to steal data so they can blackmail | I, some to steal hacking details | I and some looking to protest. They use the Internet vastly to perpetrate their dastardly act. | I During the un-targeted attacks, the hackers attack broadly. They are not focused on any particular victim, just trying out the weaknesses of infrastructure and if they see a vulnerability anywhere would take advantage of it. These type of attacks are categorised as | I Phishing | a large number of emails are sent to a large number of people with links to fake website asking for sensitive details. These are sent in the hope that some would be taken up and the links would be clicked. | Water-holing | fake websites are set up or legitimate ones compromised to confuse and take advantage of the website users. | I Targeted attacks when an hacker focuses on a particular organisation or business either to punish the organisation by a disgruntled old employee or an erate customer. The hacker must have researched on the organisation and done their background work to study the vulnerabilities | I and the attack would be tailored to achieve maximum success. Also, the attack is tailored | I specific individual, process or system in the organisation so it can be harmful. These types are | I Spear-phishing: this is when an email is targeting individuals and they are tricked into believing its legitimate. This could contain malicious links | software | that could destroy the organisation. | I DDOS: | Distributed Denial of Service | (I DDOS) | I links | botnets | deployed to slow down the productivity of systems and infrastructure, | thereby undermining the organisations | effectiveness | communication systems and processes. | Most organisations are not going to put the system of cyber mitigation and remediation in place before they get attacked. This is because it costs a lot to set up a cyber security war room and if the organisations are not attacked for probably several years, the cost and efforts appear as wasted ventures. | I Risk management teams of every organisation should carry out risk assessments of systems, | processes and the organisations internal and external environment to be able to understand if they are susceptible to attacks | I or not. | Of CRIME AND CRIMINALS | Psychology have tried since knowledge inquest | I research | I began to understand human nature and why human beings commit crime. Various theories have been derived to nationalise the cause of crimes and why criminals do what they do. | Three perspectives have been derived to explain the what contributes to crime and criminal behaviour. | I it was first perceived as an inevitable offering of urban overpopulation, poverty, bad parenting, pauperism and immigration. Much later, in the 50s, the position was that crime | I is the symptom of individual psychotic illness that would need to be addressed and treated in rehabilitation centres. | The last theory professed was the theory that whatever criminal activity prevails the society is a consequence of the social climate which is influenced by the economy of the society and that the individual has no control over it. | (Manfred, 2014) | Murry explanations, research, criminal laws, justice, imprisonment have not stopped or deterred human beings from committing crimes. | Crime is born of a desire to gain. Not that gaining is negative, but someone who subjects himself or herself to crime seeks to gain unlawfully or without too much labour. | I Criminals have I always | taken advantage of situations, circumstances, opportunities and tools to perpetrate dastardly acts. It’s sensible to say crime has been part of human nature. | I only makes sense in the natural scheme of things then that criminals would use technology to perpetrate criminal acts for gain | since it’s a tool and is readily available to all. | Nietzsche I and I Milich in their review of I James O. Wilson and Richard J. Herrnstein’s Crime and Human Nature | I corroborate this by | I seeing that the ‘I learning theory component begins with the opulent concept that behavior | I is determined by its consequences | I. | I In other words, criminal behaviour and non-criminal behaviour all have their different rewards.
Appendix E: First case of contract cheating studied

Understanding academic integrity and plagiarism in the digital age

Clare Johnson
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

system, which i aims to detect i unauthorised i modification to the data. i Availability i Availability i involves measures to make information available as and when needed to the authorized persons. Information is power, i lack of access to the right information can make a difference to individuals' course of life or an organisations success. i Attacks against availability i come in the form of i denial of service i (DoS) i attacks. i CONCLUSION i Digital forensic investigation is a crucial step in cyber security i implementation i. i Without digital forensic investigation, cyber security breaches would not be mitigated or remediated. i Any criminal investigation, evidence collection is crucial. i However, evidence collected is susceptible to modification, alteration, erasure or loss. i Due diligence must be observed in the methods used for the collection of evidence to ensure that the digital evidence is not tampered with or lost. i The examination of unallocated file space is vital during a computer forensics investigation. i When data is written to a device, data clusters are allocated to store the data. However, a file deleted by a user is not erased. This is only stored in an unallocated space until the operating system re-allocates them. This implies that a digital forensic analysis should be i consistent i of such data when collecting evidence because the unallocated data presumed deleted might be sensitive data and might need discrete handling and processing. (i (i Bassett, Bass and O'Brien 2006 p2) i) There is no going back when it comes to cyber security and digital forensics. The only way forward is to advance in our techniques and tools. The bad guys out there are constantly search, researching and discovering new ways to step their games. Organisations need to be more pro-active and be ahead of the game otherwise they will be in no time but the hackers and the overwhelming attacks against their Technology. i The strategy of governments internationally is to curb and contain crimes and cyber crimes in general. Various legislations, policies, acts, rules of laws on computer misuse and crimes are being put in place to deter men of the underworld from operations. Public bodies, private companies, charities, and all other entities are putting policies and processes in place to fight cybercrime. i UK, i The National Cyber Security Centre (NCSC) 'was established to help protect the country's critical services from cyber attacks and manage major incidents while improve the underlying security of the UK Internet through technological improvement and advice to citizens and organisations. Their vision is to help the UK the safest place to live and do business online'. i The NCSC formulates policies and act as the intelligence bureau for governmental departments by sharing and disseminating intelligence information about security breach or potential breaches. They also they are advisors for the parastatals. The formulation of all policies, strategies, legislations and all regulatory laws are vital in warding against cybercrimes. i (https://www.ncsc.gov.uk/information/about-ncsc/11.n.d.1 ; ITU, 2012, p.2) Digital Forensics require the collection of 'digital DNA' i when i a cyber crime occurs. As with criminal forensics evidence gathering, there is a need to collect the DNA evidence before the evidence is tampered with either through infiltration or loss of the evidence due to limited storage or some other limitations. Data may be lost or overwritten if due diligence is not taken care of. i Lack of industry regulations, and operating level agreements can hinder i effective cyber security or i accurate forensics of taking place. Every organisation, whether carrying out the forensics themselves or outsourcing should put in place relevant policies and clear processes and the staff should be trained so they are aware of what digital forensics entail. i A fundamental truth is that i every business or organisation should prepare for potential threat or attacks and ensure a state-of-the-art cyber provision is made before the real threat or i attack i is perpetrated i. With adequate preparation, they may be able to avert disaster and therefore i avert the need for digital forensics i investigation, which i is a i remediation i process i. When however digital forensic evidence must be gathered, it must be gathered following some basic rules; data must not be altered, reporting or documentation must be accurate, detailed, timely and meticulous, and the appropriate tool must be employed to gather the evidences. i REFERENCES i i Clancy i , i Little i (2018) I https://go.forrester.com/blog/drive-more-impact-from-data-and-analytics-with-insights-storytelling/? urn_source=forrester_promoted_articles_medium_webfeature_campaigncurrent_events dubious_content-link (Accessed: 9 January 2018) i Cyber Security Organisational i Standards i A call for views and evidence (2013) i Available at: i https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/1234465/13-659-cyber-security-organisational-standards-call-for-views-and-evidence.pdf (Accessed: 9 January 2018) i Dzoufoi i , i P and i Delfountas i (2014) Digital forensics trends and future. i International Journal of Cyber Security and Digital Forensics. The Society of Digital Information and Wireless Communications, 2,2. i http://seer.salford.ac.uk/d340/1/digital%20forensics.pdf (Accessed: 9 January 2018) i Hansen i , i Leme i and i Nissenbaum i (2009) Digital Disaster, Cyber Security, and the Copenhagen School. International Studies Quarterly 53,1155–1175 i https://www.aya.edu/projects/nissenbaum/papers/digital%20disaster.pdf (Accessed: 9 January 2018) i International Journal of Cyber Security and Digital Forensics (IJCSDF) 2(2): i 48-76 The Society of Digital Information and Wireless Communications, 2013 (ISSN: 2305-0012) i Jones, David (2016) Executive brief: closing the customer experience gap with digital-performance management i http://assets.dynamac.com/en/docs/report/trend-analysis-MTClosing-customer-experience_pdf (Accessed: 9 January 2018) i Manfredi i , i Christopher P. (2014) Crime and human nature i i A review of i Crime and Human Nature i , by James Q. Wilson and Richard J. Herrnstein. Available at: i http://www.claremont.org/erb/article/crime-and-human-nature/1 (Accessed: 9 January 2018) i NCSC i https://www.ncsc.gov.uk/information/about-ncsc/11.n.d.1 Niemet i , i Michael T. and i Milich i (1986) Book Review Crime and Human Nature: A Psychology of Criminality Vol 9.2 University of Kentucky, The i Behavior i Analyst i http://sites.tufts.edu/ncsc/news/behaviour0063-0077.pdf (Accessed: 9 January 2018) i Savito i , i Eric (2012) Gartner top 10 strategic technology trends for 2013,11 https://www.forbes.com/sites/ericavito/2012/10/23/gartner-top-10-strategic-technology-trends-for-2013/3de84d2b7615 i Schick, Shane (2015) Cybersecurity spending poised to rise in 2018, Gartner Reports. i https://securityintelligence.com/news/cybersecurity-spendingspoised-to-rise-in-2018-gartner-reports/ i Rose, Karen, Elderidge, Scott, and Chapin, Lyman (2015) The Internet of Things: An Overview: Understanding the Issues and Challenges of a More connected World. Switzerland: Internet Society, i https://www.internetsociety.org/wp-content/uploads/2017/08/IO-IoT-Overview-20151221-en.pdf i Catalog for collaboration.http://www.ncsc.gov.uk/content/files/protected_files/guidance_files/Cyber-security-risks-in-the-supply-chain_pdf i Understanding Cybercrime: phenomenon, challenges and legal response (2012) ITU i http://www.itu.int/ITU-D/bscybersecurity/legislation.html (Accessed: 9 January 2018) i Williams, Janet DAC (2012) i ACPO Good Practice Guide for Digital Evidence. Association of Chief Police Officers of England, Wales & Northern Ireland i , i http://www.digital-detective.net/digital-forensics/documents/ACPWG_Good_Practice_Guide_for_Digital_Evidence_v5.pdf
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

233

Clare Johnson


Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

The largest number of words in any one text run (excludes non-text words such as numbers) is 170

Text runs with a very large number of words would be difficult to create and could suggest pasted text.

Number of text runs with 0 words (could indicate punctuation) is 58

Text runs with no words could be due to punctuation, and can also indicate deleted text.

Number of text runs with 1 word (could indicate punctuation or spelling mistake) is 119

Text runs with 1 word are worth investigating. For example, one word changes are sometimes made to correct errors or contextual information in copied or contracted work, for example, changing the spelling from US to UK, or amending an institution name.

\texttt{w} \texttt{r} \texttt{f} \texttt{o} \texttt{n} \texttt{t} \texttt{e} \texttt{s} \texttt{c} \texttt{i} \texttt{l} \texttt{v} \texttt{a} \texttt{l} \texttt{s} \texttt{e} \texttt{v} \texttt{a} \texttt{l} \texttt{s} \texttt{e} \texttt{s} appear 1130 times inside these runs (NB currently this only collects values inside a run \texttt{w} \texttt{r} \texttt{t}):
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age
Clare Johnson
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Summary

<table>
<thead>
<tr>
<th>Item</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of paragraph elements</td>
<td>170</td>
</tr>
<tr>
<td>Total number of runs (w/r):</td>
<td>509</td>
</tr>
<tr>
<td>Total number of text nodes (w/s):</td>
<td>451</td>
</tr>
<tr>
<td>Total number of r:idR attributes (w/ridR):</td>
<td>389</td>
</tr>
<tr>
<td>Number of unique r:idR values in the document.xml file:</td>
<td>88</td>
</tr>
<tr>
<td>Total number of r:fonts (w://fonts) in entire document:</td>
<td>679</td>
</tr>
<tr>
<td>Total number of w:/14 text: elements with a value of ????????? (excludes table entries):</td>
<td>106</td>
</tr>
<tr>
<td>Total number of w:/sh: elements that are 'clear' (inside a run):</td>
<td>15</td>
</tr>
<tr>
<td>Total number of w:/fill elements (inside a run):</td>
<td>15</td>
</tr>
<tr>
<td>Total number of w:/fonts ascii counts (inside a run):</td>
<td>1130</td>
</tr>
<tr>
<td>Total number of appearances of white text:</td>
<td>0</td>
</tr>
<tr>
<td>Total number of appearances of web:</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: sh and fill may always be identical and if so shouldn't then count as duplicate flags

Flags summary

This section highlights artefacts that are considered flags for plagiarism or contract cheating. Click on the link to highlight where this occurs within the text.

<table>
<thead>
<tr>
<th>Item</th>
<th>Copy &amp; Paste Flags</th>
<th>Contract Cheating Flags</th>
<th>Line number and link</th>
</tr>
</thead>
<tbody>
<tr>
<td>w:/sh</td>
<td>15</td>
<td></td>
<td>ALL</td>
</tr>
<tr>
<td>w:/sz</td>
<td>679</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w:/fill</td>
<td>15</td>
<td></td>
<td>ALL</td>
</tr>
<tr>
<td>w:/4-text:</td>
<td>106</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White text:</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Score:</td>
<td>709</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Information extracted from core.xml

Be good to add conditional formatting to Flags column

<table>
<thead>
<tr>
<th>Category</th>
<th>Results</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Created by:</td>
<td></td>
<td>An unrecognised author should be queried</td>
</tr>
<tr>
<td>Last Modified by:</td>
<td></td>
<td>Unrecognised last modified by author should be queried</td>
</tr>
<tr>
<td>Date created:</td>
<td>2018-01-11 at 08:25:00</td>
<td>A date created value that precedes the assessment release date should be queried</td>
</tr>
</tbody>
</table>

Date last modified: 2018-01-11 at 08:44:00

236
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

<table>
<thead>
<tr>
<th>Date last modified:</th>
<th>2018-01-11 at 18:44:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference between creation time and last modification:</td>
<td>0:0:0:10:19 (Years:Months:Days:Hours:Minutes)</td>
</tr>
<tr>
<td>Number of revisions:</td>
<td>13</td>
</tr>
</tbody>
</table>

Warning - editing time indicated is less than one day

Information extracted from app.xml

<table>
<thead>
<tr>
<th>Category</th>
<th>Results</th>
<th>Flag score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word version:</td>
<td>14.0000</td>
<td></td>
</tr>
<tr>
<td>TotalTime* is:</td>
<td>605 minutes</td>
<td></td>
</tr>
<tr>
<td>Number of words given in app.xml:</td>
<td>4119</td>
<td></td>
</tr>
<tr>
<td>Number of words via text count (excludes numbers):</td>
<td>3836</td>
<td></td>
</tr>
<tr>
<td>Company:</td>
<td>Microsoft</td>
<td></td>
</tr>
<tr>
<td>The document is shared:</td>
<td>false</td>
<td></td>
</tr>
<tr>
<td>Run count:</td>
<td>509</td>
<td></td>
</tr>
<tr>
<td>Average words* per run (w/r):</td>
<td>7.54 words per run</td>
<td></td>
</tr>
<tr>
<td>Ratio of words* to total midR values (w/midR):</td>
<td>9.86 words per midR</td>
<td></td>
</tr>
</tbody>
</table>

*TotalTime represents the number of minutes that the document has been open for editing since it was created

**words value is taken via calculated count on text content as this is more accurate than app.xml value

Information extracted from settings.xml

The default language for the file is: en-GB
The midRRoot value is: 001A0A90
The midRRoot value indicates the origins of the document. Documents with matching midRRoot values originated from the same document.

This is the array of midR values appearing in the settings.xml, appearing in alpha-numerical order by hex pair

00044142 - 000456CE - 00056526 - 00060337 - 00062997 - 000691F3 - 000A0109 - 000A46EB - 001A3413 - 00216561 - 00264697 - 002A1707 - 002B2D72 - 002B313B - 002B4E70 - 002B5E4D - 002C2C26 - 002D693A - 002E6463 - 002F94AB - 002F9B34 - 00320F8E - 00323284 - 003242EB - 003544CA - 00363D90 - 00409D5C - 00444D95 - 00467778 - 00487C82 - 004C76A9 - 004D3797 - 004E1FBB - 00510CA9 - 00515752 - 0054086F - 00555148 - 00560393 - 0057404E - 00591369 - 005A0C26 - 005B17D9 - 005C3C4E - 005D019A - 005E194B - 00622A5E - 00633861 - 00679221 - 00683185 - 006C741C - 00726997 - 00726D5E - 00745B89 - 00760A95 - 0077B0DA - 00775E37 - 0077F065 - 0084468C - 008447B6 - 00867B7D - 008A34AD - 008B0809 - 008E7778 - 008F4B24 - 00914F94 - 00914F81 - 00923CA4 - 00966CD5 - 0097352D - 00973D9E - 009950F1 - 00996274 - 009A1B2C - 009B14A - 009F2746 - 009F537B - 00A256BE - 00A37341 - 00A7539E - 00A83D88 - 00AB5546 - 00AE2B68 - 00AF523E - 00BB0348 - 00BBBB43 - 00BB46C2 - 00BB5716 - 00BBB30C - 00BBB96D - 00BBFC9D - 00C71AA6 - 00C83F0D - 00C9A14D - 00C9C956 - 00D7875B - 00DA1FCA - 00DB2A2F2 - 00DB305A7 - 00E5409E - 00E54A3F - 00F516DF - 00F614CA - 00F6475C - 00FCF10D - 00FC311F - 00FC648C - 00FD14B2 - 00FD6298 - 00F06967 |

If any midR values appear in the document.xml file but not the settings.xml file they will be listed here:

No missing values

Array ()

Information extracted from styles.xml

Document statistics from styles.xml

This section of code specifies all the styles for the document, and restates the default language used

Document statistics from word/_rels/document.xml

This folder defines the relationships between images and IDs

Uncropped Image Files

Image files appear below, along with the rid value and image variable name it has been assigned. Output will also indicate if the image has been cropped.

The original of a cropped image may contain information that is useful in detection plagiarism / collusion cases

No images found

(c) 2021 No part of this code may be reproduced without express permission of the author, Clare Johnson
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Student submission output from ‘Clarify’

---

**Plagiarism and Contract Cheating Detection Tool**

This draft tool has been created as part fulfilment of a PhD in Academic Integrity at the University of South Wales.

For further details please contact Clare Johnson.

The code version is clarify/22-11-30v1.p

The file being used is files/InitialCaseStudyStudent.docx

The pdf name is files/InitialCaseStudyStudent.pdf

File type: .docx

This is a PDF version of the .docx file submitted.

You must save a PDF version of the file to the same folder as the .docx file in order for this to display the correct PDF image.

---

**Computer Forensics and crime Investigation**

[Type the document subtitle]

---

**Results Section**

**Document.xml Output**

Show or hide the document.xml code here.

Show / Hide document.xml code

The following section displays all text elements within the document. Blocks belonging to one text run are separated by a red pipe (|). Edit runs (redR) can be seen by using the toggle switches in the redR values section.

---
Appendix E: First case of contract cheating studied

Understanding academic integrity and plagiarism in the digital age

Clare Johnson

1 Cyber attacks. 2001 convention on Cyber crime 1 I In January 2014 the EU established the European Cyber Centre to support member states to respond to cyber crimes 1 Combating cyber crime will require a cohesive 1 formalized 1 international movement 1 whereby law enforcement agencies will go beyond cyber space to prosecute those engaged in illegal cyber crimes 1 Criminal techniques are constantly changing but their goals remain the same destroy or use hacking to protect a status quo 1 With information and security arrangements shared across a supply chain the cyber security 1 of any one organisation within the chain is potentially only as strong as that of the weakest member of the supply chain 1 A determined aggressor 1 notably advanced persistent threats 1 (APTs) 1 will make use of this by identifying the organisation with the weakest cyber security within the supply chain and use this vulnerability present in their systems to gain access to 1 other members of the supply chain 1 Whilst not always the case 1 it is often the smaller organisations within a supply chain who 1 due to limited resources 1 have the weakest cyber security arrangements 1 Small organisations accounted for 92 percent of the total number of cyber incidents analysed in Verizon’s 2014 Data Breach Investigation Report 4 They are often targeted because they are more vulnerable 1 represent a single point of failure 1 or have disproportionate access to important information given their size within a supply chain 1 This poses a particular risk for larger companies on whom they depend 1 The smaller firms they contract to produce the niche products required expose them to potential compromise 1 regardless of their own cyber security maturity 111 Modern Technology against Cyber Crime 1 Technology 1 the Internet and the Web from the infrastructure of the Internet of Things 1 (IoT) 1 that we have today 1 Technology serves as the hardware 1 the Things 1 while the Internet can be seen as the vehicle 1 The web is the channel through which the Internet runs and which enable communication and networking 1 These three together have transformed the different facets of human existence 1 travel 1 business 1 communication and relationship 1111 Rose 1 Eldridge 1 and Chaplin 2015 1.1111 The Internet of Things is an emerging topic of critical 1 technical 1 social 1 and economic significance 1 Consumer products 1 durable goods 1 cars and trucks 1 industrial and utility components 1 sensors 1 and other everyday objects are being combined with Internet connectivity and powerful data analytic capabilities that promise to transform the way we work 1 live 1 and play

Projects for the impact of IoT on the Internet and economy are impressive 1 with some anticipating as many as 100 billion connected IoT devices and a global economic impact of more than $1 trillion by 2025 1 At the same time however 1 the Internet of Things raises significant challenges that could stand in the way of realizing its potential benefits 1 News headlines about the hacking of Internet connected devices 1 surveillance concerns 1 and privacy fears already have captured public attention 1 Technical challenges remain and new policy 1 legal and development challenges are emerging 1 (Rose 1 Eldridge 1 Lyman Chaplin 2015 1.1111 The quest for digital transformation coupled with an array of new regulatory concerns and cyber attacks 1 is driving enterprises to invest in cyber security 1 1 According to Schick in 2017 1 enterprises worldwide spend 1 billion in 2016 in 1.1111 Key technologies here include embedded sensors 1 image recognition and NFC 1 By 2015 1 more than 70 percent of enterprises 1 a single exec will oversee all Internet connected things 1 1 The impact of these technologies on IOT can be seen in the way they are used to secure devices such as cars and buildings 1 1 Electricity supply 1 transportation infrastructure 1 military systems and logistics virtually all modern services depend on the use of ICTs 1 Cyber attacks 1 security breaches 1 can have a significant impact 1 1 Distributed Denial of Service 1 (DDoS) attacks 1 involve overwhelming the target with a flood of traffic 1 not allowing the site to execute a task before creating another request thereby overworking the server 1 The best solution is to have Intrusion detection/prevention in place such as Firewall 1 This will also prevent phishing 1 and pharming 1 Cyber Security and Crime 1 Cyber security is the protection and preservation of confidentiality 1 integrity and availability of information in the cyber space 11 Cyber Security Organisational Standards 2013 1.51 An effective cyber security reduces or mitigates the risk of cyber attacks 1 and protects an individual or organisations technology infrastructure and network 1 Cyber security as a term is often confused with the English dictionary 1 where it was used to refer to any technical glitch 1 Cyber security was first used by computer scientists in the early 1990s to underline a series of insecurities related to networked computers 1 but it moved beyond a mere technical conception of computer security when proponents argued that threats arising from digital technologies could have devastating societal effects 1 1 Psychologists have tried since knowledge 1 and moved beyond the human nature and why human beings commit crime 1 Various theories have been developed to rationalise the causes of crimes and why criminals do what they do 1 Since the beginning of the nineteenth century 1 three broad perspectives have contributed to explanations of crime 1 Early in the century 1 crime was thought to be the inevitable product of urban settings 1 poor parenting 1 pauperism 1 and the willingness of new immigrants to conform to American society 1 From the middle of the century until the 1900s 1 the dominant view was that crime is the symptom of individual psychological flaws that require probing and treatment 1 Finally 1 a theory of environmental causation emerged in which crime is attributed to social conditions whose cause is structural flaws in the economy over which ordinary individuals have no control 1 1 Psychologists have tried since knowledge 1 and moved beyond the human nature and why human beings commit crime 1 Various theories have been developed to rationalise the causes of crimes and why criminals do what they do 1 Since the beginning of the nineteenth century 1 three broad perspectives have contributed to explanations of crime 1 Early in the century 1 crime was thought to be the inevitable 1 product of urban settings 1 poor parenting 1 pauperism 1 and the unwillingness of new immigrants to conform to American society 1 From the middle of the century until the 1900s 1 the dominant view was that crime is the symptom of individual psychological flaws that require probing and treatment 1 Finally 1 a theory of environmental causation emerged in which crime is attributed to social conditions whose cause is structural flaws in the economy over which ordinary individuals have no control 1 1 Digital Forensic came about to understand 1 investigate and analyse the operations 1 patterns 1 profile of cybercrime and cybercriminals 1 Cybercrime has been the bane of business since technology’s advent 1 At the onset  there was little awareness of the cybercrime and the negative impact it had on the business world 1 As technology  knowledge sharing and networking advance  and cyber crime escalates in intensity  the awareness on the methodologies  tools and processes to combat the dark world began to grow 1 Enhancing cyber security and protecting critical information infrastructure are essential to each nation’s security and economic well being 1 1 Many examples of cybercrime 1 including the 1990s’带宽 1 and criminal laws 1 justice  imprisonment have not stopped or deterred human beings from committing crimes  Crime is born of a desire to gain  Not that gaining is negative  but someone who subjects himself or herself to crime seeks to gain unlawfully or without too much labour  Criminals have always taken advantage of situations  circumstances  tools and opportunities to perpetrate dastardly acts  It’s sensible to say crime has been part of human nature  but it only makes sense in the natural
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age  Clare Johnson

advantage of situations, circumstances, tools and opportunities to perpetrate, dastardly acts. It's sensible to say crime has been part of human nature. It only makes sense in the natural scheme of things that criminals would use technology to perpetrate criminal acts so to gain since it's a tool and is readily available to all. I Nietzel I and Milich I in their review of James Q. Wilson and Richard J. Herrnstein's Crime and Human Nature corroborate that it integrates two of America psychology's most mainstream traditions: I behaviour I learning theory and individual differences. The BOOK REVIEW 213 learning theory component begins with the operant principle that I behaviour I is determined by its II consequences. Both I criminal I and noncriminal I behaviour I have gains and losses. For example, the gains associated with I not committing a crime include having a clear conscience, avoiding I punishment I , and maintaining one's reputation. The gains associated with I criminal I behaviour include I monetary I factors, revenge, or peer approval. I Whether a crime is committed depends, in part, on the net ratio of these gains and losses for both criminal and noncriminal I behaviour I. When the ratio for committing a crime exceeds that for not doing so, there is an increased likelihood of such a crime being committed. I (11 I Nietzel I and I Milich I 1986, pp.212-213) I (4, Evaluation of current digital forensics principles: I Digital forensics is the collection, duplication, investigation, analysis and authentication of evidence when gathering and preserving evidences from computing devices in a manner that make them tamable in a court of law. This is to enable structured and legal investigation process and documentation in order to discover what happened in a cyber attack I and who was responsible for the crime. This process is very time-consuming. It requires analytical skills and investigative know-how to be able to carry it out effectively. I Digital forensics has helped in the conviction of a lot of online criminals who would otherwise have gone away with their crimes free. I The government regulations stipulates that organisations should have in place an independent audit and assurance framework to investigate for cyber security and this includes digital forensic evidence gathering policies, processes, tools and resources (Cyber Security Organisational Standards 2013, p. 6). I Digital evidence is the information found on an electronic device, that had been breached, that can be used as evidence in court. Assessing the evidence in relation to the case to be able to determine what actions to take processes digital evidence. The evidence would then be preserved with integrity and only copies of original evidences would be used for the analysis. The analysis stage is when the evidence would be extracted like recovering data from the object and interpreting it in a logical manner. Last stage is the reporting stage: when the findings are documented. (Collecting DNA evidences for forensic analysis and investigation is a great feat). I In one instance, evidences have to be collected in certain ways that makes it unsalved. I Dezfouli I , F and I Dehghantash I , A sum this up in their article published in 2014: I With the rise of challenges in the field of forensic investigations, problems that are more interesting are looming on the horizon for both victims and investigators. As computers become smaller, faster and cheaper, computers are increasingly being embedded inside other larger systems which allow information to be created, stored, processed, I analyzed I and communicated in ways that are unpredictable. I Once we gathered digital evidence from monolithic, stand-alone mainframes whereas today we have PC's, supercomputers, distributed client-server networks, laptops and smart phones, and LANs and WANs to convey information across the world, each of which is a potential source of digital evidence. Evidences stored in a computer is not unique with regard to relevancy and materiality, but because it can be easily duplicated and modified, often without leaving any traces and is readily available to a miscreant using another computer half a world away and hence, should be constrained by evolving legal standards and constraints to defend privacy issues. In general, privacy means allowing or disallowing access to information. The code of ethics requires the forensics professionals to maintain the privacy of the client. In the event of proper investigation of cases, depending on the sensitivity of the issue and the privacy of the client, there may need to be compromised. But it is also possible the victim organisation might lose out the trust over forensics team. Moreover there are organisations where in any slight leakage of the issue may attract huge media attention resulting in endangering the reputation and finally the business of organization. I In such situations, privacy rights and law enforcement's need to search and seize digital evidence during digital forensic belong together. It may also be possible that the forensics expert may not share information with any third party but takes the advantage of the confidential information of the client himself, which is also a case of violation of right to privacy. That is why, it is the policy maker's responsibility to see the impact of forensics in the broader context of business goals and make the hard decisions that trade-off forensics capabilities with issues of privacy and, correspondingly, mitigate. I Key strategies for digital forensics in order to protect privacy are selective revelation, strong audit and rule processing technologies. In the present situation, the dilemma are How to monitor digital forensics while keeping search information secret? How do we keep private information from being improperly disclosed in the name of forensics? I (1) Dezfouli I , F and I Dehghantash I , A 2014, p.48) I Digital forensics is I a new discipline I . I It has now become a crucial aspect of criminal investigation. Internet use, web and sophisticated technology are only just gaining grounds in the last few decades. Cyber security and digital forensics are in their infancies. There is so much to be learnt in this new experience that civilization has thrust on humanity. I Moreo I , compared with traditional criminal forensics, it is most of the time not as scientific or standard based as we found in the physical criminal forensic world. I Also, most organisations are just becoming aware of cyber security I and those I who are making moves to protect their network have to consider budget, human resources, space, and technical know-how. New tools (monitor of systems, monitor of monitors) have to be procured, Policies, Processes, Service Level Agreements, Operational Level Agreements have to be standardised. All these are some of the challenges evidence collection and analysis are facing. I One main issue when an incident of a major category I occurs is the time and the presence of mind to document and collate evidence for future investigation. Analysts are busy running gold command calls, mitigating the risks and threats to the network and systems, and remediating high priority incidents. I There is left little or no time to gather evidences for forensic analysis. I One big challenge can be found in I cloud space I . Data in the I cloud space I is distributed over a number of nodes, therefore breached data can affect more than one local region therefore increasing the time and fund needed to collect the evidence. I Also, there is a big challenge in locating an encrypted data in a digital crime scene environment. There is simply no time for decrypting the breached data in a critical high priority digital forensic environment. I The major challenges being faced by organisations and businesses today are the cost of maintaining the security of their systems and data. A lot goes into pre-planning, resource, tooling, and incident response. Most organisations are now outsourcing their security monitoring and handling processes to 3 rd parties to help protect their infrastructure at a considerably lower cost than it would have been if they were doing that for themselves. However this comes with its own risks too because those organisations would lose control of their process. I Reliability and completeness are two crucial needs in forensic investigation. This is broken down into three main principles. I Security assessments and policies I When carrying out security assessment, organizations must follow these steps: I (1) Strategic (Why) I involves I creating security policies, dealing with people issues, and evaluating threats and risks. I (2) (How) I involves I how the security systems are developed and implemented to meet policy requirements. I (3) Operational (What) I involves I maintaining and monitoring the enforcement of security information security policies. I SECURITY TASKS I Decisions need to be made to determine I
Operational (What) | Involves | maintaining and monitoring the enforcement of information security policies. | SECURITY TASKS | Decisions need to be made to determine: | 1 | What needs to be protected? | Who? | What are the threats and vulnerabilities? | What are the implications if they were damaged or lost? | What is the value to the organization? | Can what be done to minimize exposure to the loss or damage? | 1.5.1 Confidentiality, integrity and availability | The basic fundamental security principles are of essence when carrying out digital forensic analysis and evidence gathering. The principles are: Confidentiality | 1 | Confidentiality | Confidentiality, when referring to confidentiality in security environment, means that information that is meant to be secret should stay secret without getting into wrong hands. Information should be created, stored, communicated or transferred in manners that would help sustain its confidentiality. Sensitive information or data in the wrong hands could have serious consequences, if not fatal, to the subject or owner of the information. Although crypography and access controls are the main systems for protecting the confidentiality of information, many organisations are still behind in the usage of these tools. | 1 | Integrity | 1 | Integrity is the | maintenance of information or data in its original state. This includes the completeness and correctness of such data. Integrity involves data being maintained without modification from unauthorised persons. Integrity in cyber security can be preventive which prevents unauthorised persons to have access to modify the data or information and the detective system, which aims to detect unauthorised modifications to the data. | 1 | Availability | 1 | Availability involves measures to make information available as and when needed to the authorized persons. Information is | power | lack of access to the right information | can make a difference to | individual’s | 1 | course of life or an organisations success. Attacks against availability come in the form of denial of service (DoS) attacks. | 16. Conclusion | Digital forensic investigation is a crucial step in cyber security. Without digital forensic investigation, cyber security breaches would not be mitigated or remedied. Any criminal investigation, evidence collection is crucial. However, evidence collected is susceptible to modification, alteration, erasure or loss. | According to Bassett, Bassett and O’Brien (2006) A Computer Forensic Specialist (CFS) must follow a rigid set of methods to ensure that | computer | evidence | is | correctly obtained. These steps are outlined in Table 1, which also introduces two critical terms: unallocated file space and file slack. | The examination of unallocated file space is vital during a computer forensics investigation. When data is written to a | 1 | storage device, data clusters | from the File Allocation Table are allocated to store the data. | But when the | 1 | file is deleted | by | the | 1 | user | , the | 1 | data | is not erased. | A ‘delete’ operation will incite these data clusters to become | unallocated | , but they will still hold onto the old data until the operating system nullifies these data clusters at a later time. | The data residing in this unallocated file space can potentially contain fragments of files and subdirectories, as well as temporary files used by the application programs or operating systems. All of these types of data may contain sensitive information that can prove to be valuable during an investigation, and so it is necessary to uncover as much data from the unallocated file space as possible. Many criminals fail to recognize that the deletion process does not truly erase the sensitive data, and this is often where incriminating evidence will be discovered. | According to Bassett, Bassett and O’Brien (2006) | There is no going back when it comes to cyber security and digital forensics. The only way forward is to advance in our techniques and tools. The bad guys out there are constantly search, researching and discovering new ways to step their games. Organisations need to be more pro-active and be ahead of the game otherwise they’ll be sunk in no time but the hackers and the overwhelming attacks against their Technology. | Deterring cybercrime is an integral component of a national | 1 | cyber security | 1 | critical information infrastructure protection strategy. In particular, this includes the adoption of appropriate legislation against the misuse of ICTs for criminal or other purposes and activities intended to affect the integrity of national critical infrastructures. | At the national level, this is a shared responsibility requiring coordinated action related to prevention, preparation, response and recovery from incidents on the part of government authorities, the private sector and citizens. At the regional and international level, this entails cooperation and coordination with relevant partners. The formulation and implementation of a national framework and strategy for cyber security | thus requires a comprehensive approach. | Cyber security strategies – for example, the development of technical protection systems or the education of users to prevent them from becoming victims of cybercrime – can help to reduce the risk of cybercrime. | 40 | The development and support of cyber security strategies are a vital element in the fight against cybercrime. | (HTU, 2012, p.2) | Digital forensics require the collection of digital DNA when a cyber crime occurs. As with criminal forensics evidence gathering, there is a need to collect the DNA before the evidence is tampered with either through infiltration or loss of the evidence due to limited storage or some other limitations. Data may be lost or overwritten if due diligence is not taken care of. | 1 | Lack of industry regulations | and operating level agreements can hinder effective cyber security or accurate forensics from taking place. Every organisation, whether carrying out the forensics themselves or outsourcing should put in place relevant policies and clear processes and the staff should be trained so they are aware of what digital forensics entail. A fundamental truth is that every business or organisation should prepare for potential threat or attacks and ensure a state-of-the-art cyber provision is made before the real threat or attack is perpetrated. With adequate preparation, they may be able to avert disaster and therefore avert the need for digital forensic investigation, which is a remediation process. | It is important to note that cyber forensics evidence must be gathered, it must be gathered following some basic rules; data must not be altered, reporting or documentation must be accurate, detailed, timely and meticulous, and the appropriate tool must be employed to gather the evidences. | REFERENCES | | Clancy, L. (2018) | https://go.forrester.com/blogs/drive-more-impact-from-data-and-analytics-with-insights-storytelling?utm_source=forrester_prosolved&utm_medium=web&utm_campaign=current_events&utm_content=link | 1 | (Accessed: 9 January 2018) | Cyber Security Organisational Standards: A call for views and evidence (2013) | Available at: | | https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/324665/3-13-14-cyber-security-organisational-standards-call-for-view-and-evidence.pdf | 1 | (Accessed: 9 January 2018) | Dezhbouli, F. and Dehghantanha, A. (2012) Digital forensics trends and future. International Journal of Cyber-Security and Digital Forensics. | The Society of Digital Information and Wireless Communications, 2,2. | | http://www.sirallford.ac.uk/34014/digital%20forensics.pdf | 1 | (Accessed: 9 January 2018) | Hansen, I., Lene and A Nielsenba, | | Copenhagen School. International Studies Quarterly:53. 1155–11751 | | https://www.nyu.edu/projects/nissenbaum/papers/digital%20crime.pdf | 1 | (Accessed: 9 January 2018) | International Journal of Cyber-Security and Digital Forensics (IJCSDF) 2(2): 48–76 The Society of Digital Information and Wireless Communications, 2013 | (ISSN: 2205-0912) Jones, David (2016) Executive brief: closing the customer experience gap with digital-performance management | http://assets.dynatrace.com/midoc/report/trend-analysis-MIT-reports-customer-experience.pdf | 1 | (Accessed: 9 January 2018) | Mastrolillo, C., Christopher P. | | 2014) Crime and human nature: A review of | A Criminal and Human Nature | , by James Q. Wilson and Richard J. Herrnstein. | Available at: | http://www.claremont.edu/crb/article/crime-and-human-nature/1 | 1 | (Accessed: 9 January 2018) | Nielsen, T. Michael and I Milich, | | Richard (1986) Book Review Crime and Human Nature: A Psychology of Criminality Vol.9.2 University of Kentucky, The | Behavior Analy. 1 | file:///Users/Isoken/Downloads/bevahian00063-0077.pdf | 1 | (Accessed: 9 January 2018) | 241
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age
Clare Johnson
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

The largest number of words in any one text run (excludes non-text words such as numbers) is 199

Text runs with a very large number of words would be difficult to create and could suggest pasted text.

Number of text runs with 0 words (could indicate punctuation) is 79

Text runs with no words could be due to punctuation, and can also indicate deleted text.

Number of text runs with 1 word (could indicate punctuation or spelling mistake) is 105

Text runs with 1 word are worth investigating. For example, one word changes are sometimes made to correct errors or contextual information in copied or contracted work, for
example, changing the excellence from UK to US or remove one institution name.
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age  

Clare Johnson

Summary

<table>
<thead>
<tr>
<th>Item</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of paragraph elements:</td>
<td>181</td>
</tr>
<tr>
<td>Total number of runs (\r):</td>
<td>450</td>
</tr>
<tr>
<td>Total number of text nodes (w:\r):</td>
<td>439</td>
</tr>
<tr>
<td>Total number of xml attributes (w:xml)</td>
<td>255</td>
</tr>
<tr>
<td>Number of unique xml values in the document xml file:</td>
<td>15</td>
</tr>
<tr>
<td>Total number of fRoms (w:Fonts) in entire document:</td>
<td>365</td>
</tr>
<tr>
<td>Total number of w:14 textId elements with a value of ????????? (excludes table entries):</td>
<td>156</td>
</tr>
<tr>
<td>Total number of w:shd elements that are 'clear' (inside a run):</td>
<td>60</td>
</tr>
<tr>
<td>Total number of w:fill elements (inside a run):</td>
<td>60</td>
</tr>
<tr>
<td>Total number of w:Fonts ascii counts (inside a run):</td>
<td>139</td>
</tr>
<tr>
<td>Total number of appearances of white text:</td>
<td>0</td>
</tr>
<tr>
<td>Total number of appearances of w:shad:</td>
<td>0</td>
</tr>
</tbody>
</table>

NOTE: shd and fill may always be identical and if so shouldn't then count as duplicate flags

Flags summary

This section highlights artefacts that are considered flags for plagiarism or contract cheating

Click on the link to highlight where this occurs within the text

<table>
<thead>
<tr>
<th>Item</th>
<th>Copy &amp; Paste Flags</th>
<th>Contract Cheating Flags</th>
<th>Line number and link</th>
</tr>
</thead>
<tbody>
<tr>
<td>w:shd</td>
<td>50</td>
<td>117 118 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143</td>
<td></td>
</tr>
<tr>
<td>w:ax</td>
<td>562</td>
<td>117 118 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143</td>
<td></td>
</tr>
<tr>
<td>w:fill</td>
<td>60</td>
<td>144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166</td>
<td></td>
</tr>
<tr>
<td>w:14textId</td>
<td>156</td>
<td>160 181 182 183 216 216 217 218 219 413 414 415 416 417</td>
<td></td>
</tr>
<tr>
<td>White text</td>
<td>0</td>
<td>ALL</td>
<td></td>
</tr>
<tr>
<td>Score</td>
<td>682</td>
<td>156</td>
<td></td>
</tr>
</tbody>
</table>

244
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

Information extracted from core.xml

Information extracted from app.xml

Information extracted from settings.xml
Appendix E: First case of contract cheating studied
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Information extracted from styles.xml

Document statistics from styles.xml
This section of code specifies all the styles for the document, and restates the default language used

Show / Hide styles.xml code

Document statistics from word/_.rels/document.xml
This folder defines the relationships between images and IDs

Show / Hide rels.xml code

Uncropped Image Files
Image files appear below, along with the rid value and image variable name it has been assigned. Output will also indicate if the image has been cropped.
The original of a cropped image may contain information that is useful in detection plagiarism / collusion cases

No images found

©2021 No part of this code may be reproduced without express permission of the author, Clare Johnson
Appendix F: Researcher profile

The researcher has worked in the education sector for over 20 years. Though her first degree is in music, her academic career started as an instructor in Information Communication Technology in a community based Further Education setting, and she later managed a new programme in information security with a vocational BTEC qualification equivalent to UK A Levels. Whilst running this programme, she completed a Master’s Degree in Education and ICT, developing an interest in teaching styles and how students develop critical thinking skills. It was during the final year at this FE college that she made initial steps into studying academic integrity at PhD level. Interest in this topic began following assessment submissions by students who had plagiarised extensively, but when questioned about this plagiarism claimed that they were unaware of what they had done (poor academic practice), or indeed how and why they should reference other people’s work. Following on from FE, the researcher then moved to a Higher Education Institution as a cyber security lecturer, later becoming the academic subject manager for the department. It was during this time that the researcher first encountered academic misconduct in the form of contract cheating. Also at this time, the researcher further developed her interest in academic integrity and facilitated events for the International Day of Action Against Contract Cheating several years running.

As a cyber security expert, the researcher taught students how to carry out digital forensic investigations to a level whereby they could feasibly present an expert witness statement in court. Topics included the use of digital forensic tools such as Autopsy (free) and FTK (proprietary) as well as steganography techniques.

Whilst initially investigating the student voice in relation to academic integrity and academic misconduct, a further question arose – that of how current plagiarism detection methods work and whether they could be improved. The tools and techniques used by digital forensics teams when investigating electronic or computer crime may also be relevant in the detection of plagiarism. Indeed, they have already been used to some extent in proving theft of intellectual property albeit on a singular task at a time, so it doesn’t take a huge leap to apply this to authorship from a plagiarism detection perspective. Another area of significant interest is around the process of assignment writing. Having studied the development of critical thinking in students during her Master’s Degree, the researcher became interested in the process of writing, considering how a student goes about creating a document when they submit authentic work (i.e. work written by themselves). Would it be possible to unravel that
process, as this will look very different to work that has been copied and pasted or contracted out, and automate finding this information within a software tool? A chance encounter with a contracted author presented an opportunity to investigate this further and the author used this opportunity to analyse the extent to which the essay writing process could be extracted from the OOXML files within the submitted Word document. As a result, the first prototype tests of the methods were created, leading eventually to a web based tool. Whilst it is not expected to provide a single solution for plagiarism detection, any additional tools for assessors to use that flag unusual student behaviours when authoring assessments could be highly useful as part of a range of methods such as plagiarism text-matching software and authorship tools based on stylistics and linguistics.

During the last year of working at the institution (in the final year of the PhD research programme), the researcher was invited to take on the role of Chair of the University’s Academic Integrity Committee. This brought together academics and professional services staff from across the institution, with representation from each faculty, student services, library services and study skills teams and student casework staff. The committee aimed to explore current issues relating to academic integrity and misconduct, to summarise data on the cases of academic misconduct being reviewed, as well as discussions around penalties, support for students and staff, future challenges and more. Whilst in this role, the researcher had excellent opportunities to engage with staff across the whole organisation and within the partner colleges, helping to broaden her understanding of the key issues and challenges faced by the institution, as well as having the opportunity to shape discussions and training for staff.

At the time of writing, the author has successfully published / presented the following, all of which were subject to peer review:

**Publications:**

• Johnson, C., & Davies, R. (2020b). Plagiarism from a digital forensics perspective, In Integrity in Education for Future Happiness (pp. 78-89), Brno: Mendel University in Brno

Conference presentations:

• European Conference on Academic Integrity and Plagiarism – Full paper presented ‘Prototype tool for misconduct detection: a digital forensics approach, May 2022
• Plagiarism across Europe and Beyond – Full paper presented ‘Plagiarism from a Digital Forensics perspective’, April 2020
• Plagiarism across Europe and Beyond – Full paper presented ‘The Student Voice: What do we know about students’ perspective of Academic Integrity?‘; Working paper presented ’A practical guide to improving academic integrity in the classroom’, Ephesus, Turkey, Available at https://plagiarism.pefka.mendelu.cz/?sec=cp, May 18

Poster Presentation

• “Academic Integrity: to what extent is the student voice heard in existing research into academic integrity”, WiserD Annual Conference, Wales, July 2018
International interest shown in the tool

In addition, a number of international requests for information about the forensic techniques developed during the course of the research have been received:

- Dr. Robin Crockett, University of Northampton: discussion in relation to the unusual appearance of an unpacked student submission
- Prof. Debora Weber-Wulff (HTW Berlin University), asking if a prototype of the tool was available to review student work;
- Prof. Tara Magdalinski, (Swinburne University of Technology, Australia), asking for a prototype of the tool to review some potential contract cheating cases.
- A lecturer from a UK FE College [redacted], referred to the author by renowned expert in academic integrity Thomas Lancaster, for assistance in investigating cases of plagiarism that were not being pursued by the institution in question.
- Twitter discussion regarding RSID values and referencing my articles on this topic, October 2022 between Kane Murdoch, Cath Ellis, David Morgan linking to Johnson, Davies and Reddy (2022) and mentioning Johnson & Davies (2020a).

Citations in other publications

According to ResearchGate, the following citations have been received:


15 citations


2 citations
Contributions from other people to the research

I have been lead author on all publications, with co-author contributions detailed below.


Mike Reddy provided structural support and thematic suggestions for the article following a first draft. Ross Davies provided technical insight into the typical cyber security methods deployed in a University setting.


The opportunity to publish this paper resulted from a proposal to the European Network for Academic Integrity who were offering a fee waiver to one of their members. Mike Reddy provided structural and content advice along with proof reading support.


Ross Davies provided technical and structural support for this article.

Appendix F: Researcher profile
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Mike Reddy sparked the initial idea for this study and provided expertise in understanding students’ perceptions as well as many discussions around appropriate areas of exploration for this work.

For the research thesis in general, Mike Reddy has offered guidance, insight and expertise into many aspects of academic integrity, always challenging perceptions from both staff and students. Ross Davies provided expert guidance on the development of the software, occasionally stepping in to correct somewhat untidy coding mechanisms, and suggesting development resources to assist in the process. Susan Haywood has been a guiding light in terms of structure and overall development of the thesis.
Appendix G: Ethical approval documents

Participant information sheet – student survey

Academic Integrity, Plagiarism and Contract Cheating
Research Study

Dear Student

I would like to invite you to take part in a research study. Before you decide you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. Ask questions if anything you read is not clear or would like more information. Take time to decide whether or not to take part.

This research forms part of my PhD project, which is looking at the student voice in relation to academic integrity, plagiarism and contract cheating. My research includes a review of existing literature in the field and a critical analysis of existing information. Early findings suggest that whilst students take part in surveys on academic integrity and plagiarism, there are very few discussions directly with students. There appear to be no longitudinal studies (where students are asked their views at the start of term and again at the end of term).

Study Title: Academic integrity, plagiarism and contract cheating in Higher Education, the student perspective

Why have I been invited to take part, and what does it involve?
Several classes have been selected to take part in the research. All students in the class will be invited to complete a survey, which you will be able to do in class during October 2018. Some students may also be asked to take part in a focus group (small discussion group) or interview, which will take place outside of your classes during October 2018. The survey and focus groups will be repeated towards the end of the academic year (February 2018). This will help me to see whether your understanding and views have changed. You can take part in the survey without agreeing to take part in the focus group if you wish. Focus groups / interviews will be recorded to allow me to transcribe the discussions. After being transcribed and anonymised, the recordings will be destroyed.
Do I have to take part?
It is up to you to decide. I will describe the study and go through the information sheet, which I will give to you. I will then ask you to sign a consent form to show you agreed to take part. You are free to withdraw at any time, without giving a reason. This will not affect your studies in any way.

What sort of information do you want me to give?
Issues around academic integrity, plagiarism and contract cheating are sensitive. The purpose of the study is to gather your views, and to see how those views change during your time at the University. Support from the Student Union will be available should you need it.

What are the benefits of taking part?
I cannot promise that the study will help you, but the information I get from the study will help to increase our understanding of student perspectives in relation to academic integrity, plagiarism and contract cheating, and may help us in supporting students who need support in these areas. It may also gain an understanding of good academic practice, and why it is important to maintain academic standards.

What if there is a problem?
If you have a complaint about the research study, your experience, and / or the researcher, please contact myself in the first instance: clare.johnson@southwales.ac.uk, 01443 483246. You may also contact Mr Jonathan Sinfield, University Research Governance Officer, Research and Innovation Services, 8 Forest Grove. 01443 484518. Jonathan.sinfield@southwales.ac.uk

Confidentiality
All information provided by you will be kept anonymous and strictly confidential and no one will be able to identify you from your answers. Recordings of the focus group will be destroyed following transcription. Paper based surveys will be held in accordance with the Data Protection Act 2018:

- Individual participant surveys will be anonymous and given a research code, known only to the researcher and research supervisor
- A master list identifying participants to the research codes data will be held on a password protected computer accessed only by the researcher and supervisor
- Surveys will be stored in a locked cabinet, within locked office, accessed only by the researcher
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

- Recorded focus group / interview data will be stored on a password protected computer known only by researcher and destroyed once transcribed
- The researcher will be specifically responsible for the security of the data
- No data will be reused beyond the duration of the study
- Only the researcher and supervisor will have access to identifiable information
- Data will be anonymised prior to analysis and destroyed in accordance with UKRIO guidance (http://ukrio.org/wp-content/uploads/UKRIO-Code-of-Practice-for-Research.pdf)

What will happen if I don't carry on with the study?
You may withdraw from the study at any point and your data will be removed, up until your data is no longer identifiable.

Results of the research study?
Results and findings may be published and you may request a copy of the report at any time. No personal identifiable information will be included in the report.

Thank you for reading this information. You will be asked to sign a consent form if you are happy to take part.

Researcher
Clare Johnson, Head of Cyber Security, University of South Wales
Clare.johnson@southwales.ac.uk, 01443 483246

Supervisor
Dr Mike Reddy, Senior Lecturer, University of South Wales
Mike.reddy@southwales.ac.uk, 01443 4 82596

Contact for complaint
Mr Jonathan Sinfield, Research Governance Officer, Research and Innovation Services, 8 Forest Grove, Jonathan.sinfield@southwales.ac.uk, 01443 484518
PARTICIPANT CONSENT FORM

Full working title of research program: Academic integrity, plagiarism and contract cheating in Higher Education, the student perspective

Please tick each box if you agree with the statement.

I have read the information provided relating to the above research project in which I have been asked to participate.

The nature and purposes of the research have been explained to me, and I have had the opportunity to discuss the details and ask questions about this information.

I understand what is being proposed and the procedures in which I will be involved have been explained to me.

I understand that my involvement in this study will remain strictly confidential and that data collected will be made anonymous. I understand what will happen to the data once the research has been completed.

I hereby fully and freely consent to participation in the study.

Having given this consent, I understand that I have the right to withdraw from the program at any time without disadvantage to myself and without being obliged to give any reason.

Participant’s name (BLOCK CAPITALS)

Participant’s signature: Date:

Principal researcher’s signature: Date:

Focus Group Consent

If you are happy to be contacted to take part in the focus group, please tick this box. You can change your mind at any time.
Academic Integrity Survey

Front Sheet

This sheet will be removed prior to analysis of data

<table>
<thead>
<tr>
<th>Student Number:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Respondent Number [office use only]:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>
If you are willing to be contacted to take part in a follow up focus group please tick the box below. You can withdraw your consent at any time. Data gathered from the focus group will also be anonymised.

I am willing to be contacted to take part in a follow up focus group: 

Student Number: _________________________
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Academic Integrity Survey

By completing this survey, you are confirming that you have read and understood the information about this research project, and that you have given your consent (by signing the consent form).

Please confirm that you have signed the consent form by ticking this box □

These first few questions are about you

1. How old are you (please tick):

<table>
<thead>
<tr>
<th>17 or under</th>
<th>18-19</th>
<th>20-24</th>
<th>25-40</th>
<th>41-50</th>
<th>51 or over</th>
</tr>
</thead>
</table>

2. What is your gender (please tick):

<table>
<thead>
<tr>
<th>Female</th>
<th>Male</th>
<th>Non-binary</th>
<th>Prefer not to say</th>
<th>Other (please specify)</th>
</tr>
</thead>
</table>

3. What course are you studying (please state) eg BSc Computer Science

______________________________________

4. Where are you living?

Student Accommodation
Private rental accommodation with friends
Private rental accommodation on own / with people I don’t really know
In family home
Other (please specify)

5. If you are not living at home, is this your first time of living away from home?

Yes □ No □

6. Do you have a job which you plan to continue during your studies? If so, please state how many hours a week you work on average:

<table>
<thead>
<tr>
<th>No job</th>
<th>Less than 6 hours per week</th>
<th>6-12 hours per week</th>
<th>13-21 hours per week</th>
<th>More than 21 hours per week</th>
</tr>
</thead>
</table>

7. If you don’t yet have a job, do you plan to get one?

Yes □ No □ Not sure □
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

8. Why have you come to university (tick all that apply)?

<table>
<thead>
<tr>
<th>Option</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>To help improve my career prospects</td>
<td></td>
</tr>
<tr>
<td>To learn something new</td>
<td></td>
</tr>
<tr>
<td>To develop my skills in something I am interested in</td>
<td></td>
</tr>
<tr>
<td>Because I’m not ready for a job</td>
<td></td>
</tr>
<tr>
<td>Other (please specify)</td>
<td></td>
</tr>
</tbody>
</table>

9. Are you worried about the cost of your degree, and how you will pay it off?

- Yes
- No
- Don’t know

The next few questions will look at your awareness of academic skills and terms

10. How confident are you in your study skills (ie written skills, research skills, ability to meet deadlines)?

<table>
<thead>
<tr>
<th>Option</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Really confident</td>
<td></td>
</tr>
<tr>
<td>Mostly confident</td>
<td></td>
</tr>
<tr>
<td>Neither confident nor lacking confidence</td>
<td></td>
</tr>
<tr>
<td>Not that confident</td>
<td></td>
</tr>
<tr>
<td>Really lacking confidence</td>
<td></td>
</tr>
</tbody>
</table>

11. Do you know what plagiarism is?

- Yes
- No
- Not sure

If you answered yes, please explain briefly what you think plagiarism is:

12. Do you know what ‘academic integrity’ is?

- Yes
- No
- Not sure

If you answered yes, please explain briefly what you think academic integrity is:
13. Do you know what ‘contract cheating’ is?

Yes ☐  No ☐  Not sure ☐

If you answered yes, please explain briefly what you think contract cheating is:


14. Do you know what an essay mill is?

Yes ☐  No ☐  Not sure ☐

15. Do you think it is illegal to use an essay mill?

Yes ☐  No ☐  Not sure ☐

The next few questions refer to the last piece of written work you submitted (may be before University).
If you have never submitted written work, please go to question 20

16. Thinking about the last piece of written work you submitted, did you submit this piece of work:

<table>
<thead>
<tr>
<th>With plenty of time</th>
<th>☐</th>
</tr>
</thead>
<tbody>
<tr>
<td>Just in time</td>
<td>☐</td>
</tr>
<tr>
<td>Late, or not at all</td>
<td>☐</td>
</tr>
</tbody>
</table>

17. How confident were you with the content / subject of the assignment and your ability to answer the questions?

| Very confident | ☐ |
| Somewhat confident | ☐ |
| Neither confident nor uncertain | ☐ |
| Not very certain | ☐ |
| Very uncertain | ☐ |

18. Thinking about the support you got from your teacher / tutor on the last piece of work you submitted, did you feel:

| Very well supported by the teacher / tutor | ☐ |
| Reasonably well supported by the teacher / tutor | ☐ |
| Neither well supported nor unsupported by the teacher / tutor | ☐ |
| Not very well supported by the teacher / tutor | ☐ |
| Poorly supported by the teacher / tutor | ☐ |
19. Overall, when you think about this piece of work, did you feel any of the following? (tick all that apply)

- I felt stressed about doing the work
- I was worried about getting a good mark
- I struggled to get the work done in time
- I didn’t feel like I had enough support to complete the work
- I didn’t feel any of the above, and was happy doing the work
- Other (please specify)

The next few questions will ask you to think about your own views and experiences

20. During your previous education (e.g. BTEC / A levels or similar) have you ever done the following? (tick all that apply)

- Copied from a friend?
- Copied a sentence from the Internet straight into your work without referencing it (saying where it came from)?
- Copied a paragraph or more from the Internet straight into your work without referencing it?
- Asked someone else to write part of your assignment / test for you?

21. Do you know of anyone who has ever cheated on an assignment, by copying text from someone else or from the Internet / printed materials, or by asking someone to do the work for them, regardless of whether that is for money or not?

- Yes, I know someone who has copied text
- Yes, I know someone who has asked someone to do the work for them
- If so, was this for money, other incentive, for free?
- No, I don’t know anyone

If you answered No above, please go to question 25

22. If yes, how did you feel about it? (tick all that apply)

- It made me angry
- I reported them
- I wanted to report them but didn’t
- I didn’t really mind
- I don’t know
- Other (please specify)
23. Did this person get caught?

Yes ☐ No ☐ Not sure ☐

24. If yes, what punishment did they receive?

<table>
<thead>
<tr>
<th>Punishment</th>
<th>☐</th>
<th>☐</th>
<th>☐</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade for the assessment was cancelled</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Grade for the module was cancelled</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>All grades for the year were cancelled and student required to resit</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Student was excluded from further study</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>They were caught but no punishment was given</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Don't know</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

25. In which situation do you think it would be acceptable to get someone else to write an assignment for you, or copy the majority of the text from the Internet / printed materials? (tick all that apply)

<table>
<thead>
<tr>
<th>Situation</th>
<th>☐</th>
<th>☐</th>
<th>☐</th>
</tr>
</thead>
<tbody>
<tr>
<td>If you run out of time to do it yourself</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If you don’t think you are going to pass</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If you are feeling very stressed</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If you have run out of time to do it yourself</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>If you think the assignment is pointless</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Never</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>Other (please specify)</td>
<td>☐</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

26. Have the penalties for academic misconduct been explained to you?

Yes ☐ No ☐ Not sure ☐

27. Do you know what to do if you know you are going to be unable to complete an assignment on time?

Yes ☐ No ☐ Not sure ☐

28. Do you think you might ever find yourself in a position where you might copy an assignment from someone else or the Internet / printed materials, or ask someone to write it for you?

Yes ☐ No ☐ Not sure ☐
29. Please write below any thoughts you might have on why people might consider breaking the rules of academic integrity when it comes to written assignments:


30. Is there anything else you would like to add?


Thank you for taking the time to complete this survey. Your answers will be treated in the strictest confidence.

If you have any questions about the survey, please contact clare.johnson@southwales.ac.uk, 01443 483246
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

Ethics forms for contract cheating detection work

University of South Wales
Proforma for Ethical Appraisal of Research Proposals

Before completing this form please read the University’s handbook General Ethical Guidance for Research and Consultancy for guidance, in addition to any specialist guidance for relevant disciplines. Please ensure you are aware if the work proposed in this document requires a standard ethics review, or a light touch review. Please see appendix A. Please attach any documents required in the recruiting of human participants, such as information sheets and consent forms.

Please contact you Faculty Ethics Champion with any queries, or request for document templates.

Section A (i.e. this page) is to be completed by all proposers. Section B of this form is to be completed only for proposals in which ethical issues have been identified.

Section A:

Name and department of proposer: Clare Johnson, FCES

Type of work covered by this form:

(tick one as relevant)

X□ MPhil/PhD research project
□ Postdoc research project
□ Staff research project
□ Other (please explain)

This work requires: □ Light touch
X□ Standard review

………………………………………………………………………

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

Project reference number: ………………………………………

Please attach a copy of the proposal descriptor as prepared for validation or registration and complete and attach the following as necessary:

□ No ethical issues have been identified as arising from this proposal.

□ Attached are the ethical issues which have been identified as arising from this proposal and a statement of how they have been, or will be addressed (please proceed to complete Sections B).

□ Advice is sought on the ethical issues noted on the attached.

(Please proceed to complete Sections B)

Signed by proposer … C S Johnson ………………… Date 19 March 2019
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

FEC use:

Ethical Opinion:

☐ ☐ ☐

Favourable Unfavourable Ref’d to ESG

Signature of FEC ……………………………… Date ………………………………
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

Section B:
Name and Department of proposer: Clare Johnson FCES
Names and Departments of co-investigators and collaborators (if any):

Dr Ross Davies, FCES

What are the objectives of the proposed research?

(Please use language comprehensible to a lay person.)

This research breaks new ground in analysing the metadata of an assignment known to have been purchased (contract cheating), and using that metadata to establish a pattern of editing that could be used to ascertain future instances of contract cheating.

The findings will be presented at a conference in June, and will be submitted to an Academic Integrity journal for publishing.

Why is this research area important and what is the scientific justification for the research?

Contract cheating is an increasing problem in academia. Educational awards are given on the strength of work submitted: if that work has been prepared by someone else, and not the student submitting it then there is a danger of granting high level awards (degrees, postgraduate degrees) to students who may not be capable or skilled at that level. Tools are available to help detect plagiarism (copy and paste) and some work has been done on linguistics and style to help identify contract cheating, but no work has been found during the literature review on using digital forensics techniques to identify such contracted work.

Where will the research take place? At the University of South Wales
Will the research involve anyone from the following groups: No

- Children under 16
- Adults who are unable to give consent
- Prisoners
- Healthy Volunteers
- Adults with learning disabilities
- Adults who have a terminal illness
- Young Offenders
- Other vulnerable groups

Will the research involve any of the following activities: No

- Examination of medical records by those who would not normally have access to them
- Sharing of data with other organisations
- Export of data outside the European Union
- Use or storage of personal addresses, postcodes, faxes, e-mails or telephone numbers
- Publication of data that might allow identification of individuals?
- Publication of data that might allow identification of postcode level or smaller geographical areas?

Give a summary of the design and methodology of the research including the number of participants to be recruited and how the number was decided upon (if any)

Methodology

This research breaks new ground in analysing the metadata of an assignment known to have been purchased (contract cheating), and using that metadata to establish a pattern of editing that could be used to ascertain future instances of contract cheating.

The findings will be presented at a conference in June, and will be submitted to an Academic Integrity journal for publishing.

Contract cheating is an increasing problem in academia. Educational awards are given on the strength of work submitted: if that work has been prepared by someone else, and not the student submitting it then there is a danger of granting high level awards (degrees, postgraduate degrees) to students who may not be capable or skilled at that level. Tools are available to help detect plagiarism (copy and paste) and some work has been done on linguistics and style to help identify contract cheating, but no work has been found during the literature review on using digital forensics techniques to identify such contracted work.
An allegation from a contract author was received suggesting that the contract author had written an assignment for a student. Copies of the work written by the contract author, along with contract details, emails etc were submitted as evidence. The work submitted by the student was compared with that of the contract author and deemed to have originated from the contract author. Standard procedures for academic misconduct were followed, and at the academic misconduct hearing, the student confessed to having contracted the assignment out.

As a known piece of contracted work, analysis using standard digital forensic techniques will be carried out, including reviewing easily found metadata attached to the submitted work (date modified, author, editing time etc). A further analysis will then be carried out by extracting the sub content of the document submitted. This will be coded to establish the type of editing carried out, how much editing took place and how this compares to the work sent in by the contract author.

How will potential participants in the study be (i) identified, (ii) approached and (iii) recruited (if any)?

As a case study, one piece of work will be used for analysis. However, it would be beneficial to locate additional examples of contracted work to add supporting evidence. Student Casework will be contacted to see if one or two other examples are available to add to the study.

Please state the main ethical issues with the proposed research, detailing any potential for adverse effects, risks or hazards, pain, discomfort, distress, or inconvenience to the research participants or researchers themselves.

Academic misconduct is an important subject and can have significant implications on a student’s progression. This research will be used in the first instance to establish whether analysis of metadata can be reliably used to indicate contract cheating. At this stage the research will not be used to determine contracted work, which will be identified through the normal procedures of marking / moderation.

How will these issues be addressed?

At this time, the research will only be carried out on work that is known to be contracted – ie has gone through an academic misconduct panel and already been confirmed as contracted work. Therefore the research is proof of concept at this stage, and the student(s) concerned will have already been through the misconduct procedure. Work will be anonymised and sufficiently redacted to prevent any identification of individual or modules, though it is probable that the area of study may be visible (as snippets of text may be used as supporting evidence).

Have ethical issues of confidentiality and consent been considered? Yes X No N/A

Does the Chief Investigator or any other investigator/collaborator have any direct personal involvement (e.g. financial, share-holding, personal relationship etc.) in the organisations sponsoring or funding the research that may give rise to a possible conflict of interest? Yes / X No. (Please give details below if you answered Yes)

Please indicate if there are any other factors which might call into question the objectivity of the research: NA

Indicate any issues on which you would welcome ethical advice

Appendix A

It is likely that a Standard Ethics Review of your proposal will be required and this will apply to most research projects where human subjects are involved.
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

In order for an application to qualify for ‘light touch’ scrutiny, the research must not fall into any of the following categories (which replicate research described in the ESRC Research Governance Framework as involving more than minimal risk):

- Research involving vulnerable groups – for example, children and young people, those with a learning disability or cognitive impairment or individuals in a dependent or unequal relationship.
- Research involving sensitive topics – for example, participants’ sexual behaviour, their illegal or political behaviour, their experience of violence, their abuse or exploitation, their mental health or their gender or ethnic status.
- Research involving groups where permission of a gatekeeper is normally required for initial access to members – for example, ethnic or cultural groups, native peoples or indigenous communities.
- Research involving deception or which is conducted without participants’ full and informed consent at the time the study is carried out.
- Research involving access to records of personal or confidential information, including genetic or other biological information concerning identifiable individuals.
- Research which would induce psychological stress, anxiety or humiliation or cause more than minimal pain.
- Research involving intrusive interventions – for example, the administration of drugs or other substances, vigorous physical exercise or techniques such as hypnotherapy. Participants would not encounter such interventions, which may cause them to reveal information which causes concern, in the course of their everyday life.

Low risk research should, therefore, be characterised by the absence of all of the above components. It should be noted that no category of research (e.g. undergraduate research dissertations) will always meet the low risk criteria.
Research Ethics Form

- This form should be completed for all projects seeking research ethics approval.
- This form should be completed in line with the relevant guidelines appropriate to your area of study.
- Research studies that need research ethics approval must not commence until ethical approval has been granted; it is the responsibility of the Researcher for ensuring due diligence and accountable decision making.
- Studies that intend to run within the NHS, Ministry of Defence, or the Prison Service will require additional approval – please seek advice from your Faculty Ethics Champion or the Research Governance Officer BEFORE completing this application form. Separate or replacement ethics forms may be required.
- Please familiarise yourself with the criteria for **HIGH** and **LOW** risk research.

<table>
<thead>
<tr>
<th>Overview of Proposed Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Title:</td>
</tr>
<tr>
<td>This study is:</td>
</tr>
<tr>
<td>Name of USW Principal Investigator:</td>
</tr>
<tr>
<td>Is this a student project?</td>
</tr>
<tr>
<td>Please list all co-investigators involved in the project</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>Dr Ross Davies</td>
</tr>
<tr>
<td>Is this project funded?</td>
</tr>
<tr>
<td>Click here to enter text.</td>
</tr>
</tbody>
</table>
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

<table>
<thead>
<tr>
<th>Intended start date:</th>
<th>19/03/2019</th>
<th>End date:</th>
<th>01/12/2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please list any codes of conduct that relate to your discipline or area of study?</td>
<td>Student Code of Conduct; Academic Misconduct Regulations</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Briefly outline of your Research Study (max 1000 words) to include:

Rationale, research aims and/or questions, and overview of methodology

This research breaks new ground in using digital forensics techniques to analyse the metadata of an assignment known to have been purchased (contract cheating), and using that metadata to establish a pattern of editing that could be used to ascertain future instances of contract cheating.

The research question is “Can digital forensics techniques used in data recovery and crime investigation be used to identify contract cheating?”

Contract cheating is an increasing problem in academia. Educational awards are given on the strength of work submitted: if that work has been prepared by someone else, and not the student submitting it then there is a danger of granting high level awards (degrees, postgraduate degrees) to students who may not be capable or skilled at that level.

There is considerable existing research into contract cheating. The QAA published a document in 2017 ‘Contracting to Cheat in Higher Education’ discussing the extent of the problem and some ideas for mitigating it, and identified 17,000 academic offences in 2016 (https://www.bbc.co.uk/news/education-47629043). Thomas Lancaster has written many articles on the issue, including a focus on essay mills, discussions on blackmailing students who have contracted work and how ‘student cheating is damaging education’. Phil Newton of Swansea University is also very active in this field and his research suggests that contract cheating is on the increase (https://www.swansea.ac.uk/press-office/latest-research/lateststudyrevealssharpriseinessaycheatinggloballywithmillionsofstudentsinvolved.php).

Recently, the investigator of this study received direct correspondence from a contract author asking for their services to be added to the University’s Referencing and Good Academic Practice pages, with follow up emails to chase a response. Furthermore, in early January 2019, homes in Treforest received a leaflet advertising essay writing services — clearly targeting USW students — saying ‘Any Academic Level’, ‘Any Subject Area’, local help, ‘foreign student friendly’ and even offering a financial incentive for referring a friend. WonkHE reported on the threat of essay mills in 2018: https://wonkhe.com/blogs/universities-to-blame-for-profiteering-essay-mills/ In Australia and Ireland, steps have been taken to make it illegal to advertise essay writing services.

Tools are available to help detect plagiarism (copy and paste) and some work has been done on linguistics and style to help identify contract cheating, but no work has been found during the literature review on using digital forensics techniques to identify such contracted work.

Methodology:

An allegation from a contract author was received suggesting that the contract author had written an assignment for a student. Copies of the work written by the contract author, along with contract details, emails etc were submitted as evidence. The work submitted by the student was compared with that of the contract author and deemed to have originated from the contract author. Standard procedures for academic misconduct were followed, and at the academic misconduct hearing, the student confessed to having contracted the assignment out.
As a known piece of contracted work, analysis using standard digital forensic techniques will be carried out, including reviewing easily found metadata attached to the submitted work (date modified, author, editing time etc). A further analysis will then be carried out by extracting the sub content of the document submitted. This will be coded to establish the type of editing carried out, how much editing took place and how this compares to the work sent in by the contract author.

The findings will be presented at a conference in June, and will be submitted to the partner journal for the conference - Journal of Academic Ethics (Springer) - for publication.

There is understandably a concern that disclosing an incident of contract cheating has occurred within the University could bring USW into disrepute. However, given that the instance was detected, and the student punished in accordance with university regulations, we believe that this is positive for the University. In any case, any features identifying the student or the University will be removed, and it will not be possible to ascertain where the case in question originated.

<table>
<thead>
<tr>
<th>Research setting. Please state locations of data collection.</th>
</tr>
</thead>
<tbody>
<tr>
<td>University of South Wales</td>
</tr>
<tr>
<td>Your relationship to participants (if any)</td>
</tr>
<tr>
<td>None specific, though the participant may be a student on one of our courses</td>
</tr>
<tr>
<td>Describe the sample and sampling strategy, including who your intended sample are. Please describe the number of participants you require.</td>
</tr>
<tr>
<td>An assignment which has been proven as contracted out will be used for the case study. The student submitting the assignment has attended an Academic Misconduct hearing in the normal manner and at this panel confirmed that the work submitted was written by a contracted author. It will be beneficial to locate other contracted work to support the initial case study. If approved, this will be done via the Student Casework department.</td>
</tr>
<tr>
<td>How will data be collected?</td>
</tr>
<tr>
<td>The initial case study work has been submitted to the department. The contract author has sent in their work via email. Other contracted examples will be gathered from Student Casework.</td>
</tr>
<tr>
<td>How will data be analysed?</td>
</tr>
<tr>
<td>Digital Forensic techniques will be used to analyse the work. This will be done by reviewing the sub documents that are packaged within the submitted Word document, that would not normally be accessed or visible. Tags and metadata will be reviewed and encoded to establish patterns and themes within the document which can be compared to a known original submission to see where differences and / or anomalies exist.</td>
</tr>
<tr>
<td>Informed Consent</td>
</tr>
<tr>
<td>Describe how, and from where you will gain access to the participants?</td>
</tr>
<tr>
<td>Student Casework. As all data will be anonymised and redacted to ensure identification of specific modules or students is not possible, it is not deemed necessary to gain consent from the students, who are not active participants in the process.</td>
</tr>
</tbody>
</table>
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

<table>
<thead>
<tr>
<th>Please include the name and contact details of any ‘gatekeepers’ of your intended research participants.</th>
<th>Only the two authors will have access to any original materials.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>How will you provide participants with information about the study before they agree to participate?</th>
<th>We do not anticipate the need to inform those students whose work we analyse as they are not active participants.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Will you inform participants of a complaints procedure?</th>
<th>NA</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>How will you record informed consent from the participants?</th>
<th>NA</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Checklist for Managing Informed Consent</th>
<th></th>
</tr>
</thead>
</table>

| All participants will be given an appropriate level of information about the project and be given adequate time to think about the information before being asked to agree to participate | NA ☒ YES □ |

| All participants will be informed of any limits to confidentiality, or circumstances where confidentiality may need to be broken, such as the disclosure of criminal activity, mal practice, misconduct or risk of harm. | NA ☒ YES □ |

| All participants being asked to provide personal data will have the following statement on the consent form or their questionnaire ‘I consent to the processing of my personal information for the purposes of this research study. I understand that such information will be treated as confidential and handled in accordance with the General Data Protection Regulations 2016 and will be destroyed once it is no longer required’ | NA ☒ YES □ |

| All respondents will be told that they can withdraw at any time, and ask for their data to be removed from the project, until it is no longer practical to do so (e.g. when a report or publication has been written and submitted) | NA ☒ YES □ |

| All participants taking part in an interview, focus group, observation (or other activity which is not questionnaire based) will be asked to sign a consent form | NA ☒ YES □ |

<table>
<thead>
<tr>
<th>All participants completing a questionnaire will be informed:</th>
<th></th>
</tr>
</thead>
</table>

| i. On the Information Sheet that returning the completed questionnaire implies consent to participate | NA ☒ YES □ |

| ii. At the beginning and end of the questionnaire ‘that returning the completed questionnaire implies consent to participate’. Conditions of consent must also be provided such as those found on a consent form | NA ☒ YES □ |
### Appendix G: Ethical approval documents

Understanding academic integrity and plagiarism in the digital age

Clare Johnson

<table>
<thead>
<tr>
<th><strong>Duty of Care</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Is there any chance that a participant’s well-being might be affected by this research? If YES, how will this be managed? What provision will be put in place by the researcher?</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>Are there any risks to the participants? Please identify and describe how they will be managed/mitigated.</strong></td>
<td>No – any instances of contract cheating will be identified through the usual marking and moderation process, and findings from this study will not, at this time, impact on the student submitting the work in any way.</td>
</tr>
<tr>
<td><strong>Does any of the research team have any potential conflict of interest with the study population or study funders? If so, please list them and describe how the conflicts will be mitigated.</strong></td>
<td>NA</td>
</tr>
<tr>
<td><strong>Are there any issues around researcher safety and if so how will risks be managed?</strong></td>
<td>NA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Managing Data</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Who will have access to the raw data?</strong></td>
<td>Only the researchers (Clare Johnson and Ross Davies)</td>
</tr>
<tr>
<td><strong>Will data be passed outside of the university network?</strong></td>
<td><strong>NO ☒</strong> If YES, do you confirm personal data will always be secured by encryption or password protection? <strong>YES ☐</strong> <strong>NA ☒</strong></td>
</tr>
<tr>
<td><strong>How will you ensure the confidentiality of the research participants?</strong></td>
<td>All data will be anonymised and redacted to ensure confidentiality is maintained.</td>
</tr>
<tr>
<td><strong>If confidentiality is limited please describe how this will be managed. Under what circumstances would confidentiality need to be broken?</strong></td>
<td>NA</td>
</tr>
<tr>
<td><strong>Will research findings be fed back to the research participants? If YES, how?</strong></td>
<td>NA</td>
</tr>
<tr>
<td><strong>Will the research be disseminated to the wider community? If YES, how?</strong></td>
<td>Yes, through a conference paper and subsequent journal article.</td>
</tr>
<tr>
<td><strong>How will you ensure the anonymity of the research participants?</strong></td>
<td>No identifying features will be used in the conference presentation nor journal article.</td>
</tr>
</tbody>
</table>
### Checklist for Managing Issues of Confidentiality and Anonymity

<table>
<thead>
<tr>
<th>Statement</th>
<th>NA</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Questionnaires will be returned anonymously and indirectly (not handed to the researcher)</td>
<td>☒</td>
<td>☐</td>
</tr>
<tr>
<td>Questionnaires and/or interview transcripts will only be identifiable by a unique identifier (e.g. code/pseudonym)</td>
<td>☒</td>
<td>☐</td>
</tr>
<tr>
<td>Lists of identity numbers or pseudonyms linked to names and/or addresses will be stored securely and separately from the research data</td>
<td>☐</td>
<td>☒</td>
</tr>
<tr>
<td>All name of people, places or organisations which could lead to the identification of individuals or organisations will be changed</td>
<td>☐</td>
<td>☒</td>
</tr>
<tr>
<td>I confirm that my research records will be held securely at USW according to the General Data Protection Regulations that came in force in 2018 and in accordance with USW research data management guidelines</td>
<td>☐</td>
<td>☒</td>
</tr>
<tr>
<td>I confirm that I will not use the research data for any other purpose than that declared on the study consent form</td>
<td>☐</td>
<td>☒</td>
</tr>
<tr>
<td>Data will be stored on a password protected personal computer. Documents will be kept locked in a secure filing cabinet or similar. Data will be destroyed in line with the data Protection Act and preserved/destroyed in line with the USW data management guidance</td>
<td>☐</td>
<td>☒</td>
</tr>
</tbody>
</table>

### Insurance / Indemnity

<table>
<thead>
<tr>
<th>Statement</th>
<th>NA</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Please confirm if the USW is acting as scientific sponsor for this study - where USW is taking responsibility for the design of the study, the soundness of the science, the project safety, and the ability for the project to achieve its aims?</td>
<td>☒</td>
<td>☐</td>
</tr>
</tbody>
</table>

Please indicate all that apply: NA to all ☒

- Investigating or participating in methods of contraception? YES ☐
- Assisting with or altering the process of conception? YES ☐
- The use of drugs? YES ☐
- The use of surgery? (other than biopsy) YES ☐
- Genetic engineering? YES ☐
- Participants under 5 years of age? (other than activities above) YES ☐
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

- Participants known to be pregnant? (other than activities above) | YES □
- Pharmaceutical/medical product/appliance designed or manufactured by host institution? | YES □
- Work outside of United Kingdom? | YES □
- Other approvals, for example approval of external organisations allowing you access to their participants. | YES □

If any of the above apply, YOU will need to notify the University Research Governance Officer before you start your research project.

<table>
<thead>
<tr>
<th>Security Sensitive Material</th>
</tr>
</thead>
<tbody>
<tr>
<td>Will your project involve any of these?</td>
</tr>
<tr>
<td>Ministry of Defence-commissioned work on military equipment or policy</td>
</tr>
<tr>
<td>EU security research including policy development</td>
</tr>
<tr>
<td>All work related to extremist groups (e.g. related to animal rights campaigners)</td>
</tr>
<tr>
<td>IT encryption design for public bodies or businesses</td>
</tr>
<tr>
<td>All work related to terrorism</td>
</tr>
</tbody>
</table>

If you respond YES to any of the above please ensure your research is also registered via the USW PREVENT for RESEARCH.

<table>
<thead>
<tr>
<th>Attachments to this Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>As part of an ethics review it is necessary to submit copies of additional documents. Please indicate the documents included to accompany this ethics application form.</td>
</tr>
<tr>
<td>Participant Information Sheet</td>
</tr>
<tr>
<td>Consent Form</td>
</tr>
<tr>
<td>Data collection tools (questionnaire, interview questions, survey, etc.)</td>
</tr>
<tr>
<td>Researcher Safety Protocol / risk assessment</td>
</tr>
<tr>
<td>Letters/confirmation from gatekeepers granting access to participants</td>
</tr>
</tbody>
</table>
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age
Clare Johnson

<table>
<thead>
<tr>
<th>Other supporting documents such as debrief sheets, information about support networks, sources of counselling, etc. Please list those that are attached.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA ☒ YES ☐</td>
</tr>
</tbody>
</table>

Click or tap here to enter text.

**Applicant’s Declaration**

<table>
<thead>
<tr>
<th>If your project is approved you must follow the research protocol and documents that have been approved. If your application is not approved you will need to refer to this version of your application when preparing your re-submission. Please note if you intend on deviating from the approved protocol or documentation you will need to request approval for any changes.</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☒</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I understand that failure to follow my approved protocol constitutes research misconduct and the policy for such offences will be followed in such an instance. Please see the USW Research Misconduct Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☒</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I understand that I must contact the relevant Faculty Research Ethics Chair and seek approval for any amendments to, or deviations from, the approved project protocol before continuing with my research project</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☒</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I have read and agree to abide by the guidelines set out with in the USW Research Ethics Policy</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☒</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I confirm that all procedures that will occur within the research will adhere to USW Policy on Health and Safety</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES ☒</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Print name:</th>
<th>Date: 19/03/2019</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clare Johnson</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Signature: C S Johnson</th>
</tr>
</thead>
</table>

**Decision**

<table>
<thead>
<tr>
<th>Approval</th>
</tr>
</thead>
</table>

277
<table>
<thead>
<tr>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Further Information needed and Resubmission required</td>
</tr>
<tr>
<td>Rejected</td>
</tr>
<tr>
<td>ethicsform</td>
</tr>
</tbody>
</table>
Analysis with digital forensics tool

INFORMATION SHEET GUIDANCE & TEMPLATE

Creating a tool to detect forensic artefacts that could indicate plagiarism or contract cheating

I would like to invite you to take part in a research study about plagiarism and contract cheating. Before you decide you need to understand why the research is being done and what it would involve for you. Please take time to read the following information carefully. Ask questions if anything you read is not clear or would like more information. Take time to decide whether or not to take part.

I have been developing an automated tool to analyse information in documents that is usually hidden. This information can tell me how a document has been put together, and I think this will help assessors to detect when work has not been genuinely written by the student. The tool will highlight this information and notify the assessor that the work shows flags that may be associated with copy and paste plagiarism and contract cheating.

What is the purpose of the study?

Academic misconduct, such as plagiarism and contract cheating, is a huge challenge for academia. Assessors have tools like Turnitin to help detect text matches, but very few tools exist that analyse a document in more detail, analysing the information behind the document and the way it has been put together. This tool will examine what is going on behind the scenes, highlighting potential flags for academic misconduct. Developing this may help alert assessors to potential academic misconduct, which in turn helps to protect the value of the qualification being awarded, and may assist in ensuring that students achieve their awards on the basis of their own work, and not that of others.

Why have I been invited?

You will be invited to take part if you are comfortable that work you have previously submitted for assessment is your own work, and that it has not been plagiarised nor contracted out. You should only submit work that you know is genuinely yours. Work containing some information copied from the internet and cited correctly is expected as this is part of normal academic practice. It is important that the tool can work out what threshold is acceptable for correctly cited work. Any written assessment is suitable, provided it is available in Word format. You should NOT volunteer to submit work that you know to be plagiarised or that you have purchased or obtained from someone else.
Do I have to take part?

It is entirely up to you to decide whether to take part or not. I will describe the study and go through the information sheet, which I will give to you, and will answer any questions you might have about the research and your involvement. If you are willing to take part, I will then ask you to sign a consent form to show you agreed to take part. You are free to withdraw your work until the point at which it is anonymised and aggregated with the data set, at which point it will not be possible to extract your work.

What happens if I agree to take part?

I will ask you to send me a copy of your assignment. It must be the original work submitted for assessment. If you prefer, I can download the assignment myself from Blackboard and if so, I will need your student number, course and module title, as well as the name of the assignment submitted. Your work will then be anonymised with a simple title (e.g. L5100) and analysed using the tool. Data extracted from the analysis will be compiled, and at this stage the data will not be identifiable to anyone else.

Expenses and payments

Unfortunately I am not able to provide any payment for your participation.

What will I have to do?

Once you have submitted your document to me (or sent me details of its location on Blackboard) there is nothing further for you to do. You will not be contacted again, and any findings from your paper will form part of the research data set, fully anonymised.

What are the possible disadvantages and risks of taking part?

At this stage the tool is being developed, and no thresholds are yet available to determine whether a document is likely plagiarised / contracted or genuine, so there is no risk to you in submitting your work. The aggregated and anonymised data will help to inform future development for the accuracy of the tool, and findings at this stage will not be used as part of the assessment process.

What are the possible benefits of taking part?

I cannot promise that the study will help you directly, but if successful, this tool may help to detect plagiarism and contract cheating in future student work, which helps to protect the integrity of the qualifications you are working towards.

What if there is a problem?

If you have any complaints about the study, or your experience during the study, you can contact the researcher:

Clare.johnson@southwales.ac.uk

If you wish to make a formal complaint about the research, you can contact the University’s Research Governance Manager:

Jonathan.sinfield@southwales.ac.uk (01443 484518)
Appendix G: Ethical approval documents

Understanding academic integrity and plagiarism in the digital age

Clare Johnson

Data Protection Privacy Notice

The data controller for this project will be the University of South Wales. The University Compliance Manager provides oversight of university activities involving the processing of personal data. The University of South Wales Compliance Manager is Mr Rhys Davies (rhys.davies@southwales.ac.uk).

Your personal data will be processed for the purposes outlined in this information Sheet. Standard ethical procedures will involve you providing your consent to participate in this study by completing the consent form that has been provided to you. However, the legal basis on which this task is being performed is public interest, approved by the Faculty Research Ethics Committee.

If you are concerned about how your personal data is being processed, please contact Compliance Manager, Mr Rhys Davies (rhys.davies@southwales.ac.uk).

Details of your individual rights are available on the ICO website at: https://ico.org.uk/fororganisations/data-protection-reform/overview-of-the-gdpr/individuals-rights/

Will my taking part in the study be kept confidential?

All data collected during the study will be kept strictly confidential, and any information about you that leaves the university will have no identifying features so that you cannot be recognised.

- Data will be collected via email and possibly directly from Blackboard, for students who have volunteered.
- Electronic data that identifies participants will be held on a password protected computer, and will be renamed to anonymise the file for analysis. Once analysis has been completed, the data will be aggregated and anonymised for reporting findings.
- All data will be treated in the strictest confidence.
- Aggregated and anonymised data may be used for research reports.
- Identifying data will be destroyed once the aggregated data has been collated in line with UKRIO guidance.
- There will be limits to confidentiality if analysis of the work submitted strongly suggests misconduct has occurred. To this end you are advised not to submit work that you know to be plagiarised or contracted out. In this case, findings will need to be discussed with the module leader.
- Work submitted will only be used for this specific study.
Appendix G: Ethical approval documents
Understanding academic integrity and plagiarism in the digital age

Title of Project: **Creating a tool to detect forensic artefacts that could indicate plagiarism or contract cheating**

Name of Researcher: **Clare Johnson**

Name of supervisor: **Dr Ross Davies**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>I confirm that I have read and understand the information sheet dated 27/04/21 (version 1.0) for the above study. I have had the opportunity to consider the information, ask questions and have had these answered satisfactorily.</td>
</tr>
<tr>
<td>2.</td>
<td>I understand that my participation is voluntary and that I am free to withdraw up until the point that is anonymised and aggregated with the data set, without giving any reason, without any consequence to myself.</td>
</tr>
<tr>
<td>3.</td>
<td>I agree to my written work being included in this study and it’s been explained how this data will be stored, destroyed, anonymised. Who will have access to it, and how long it will be kept.</td>
</tr>
<tr>
<td>4.</td>
<td>I confirm that the work I am submitting to the study has not been plagiarism or contracted out, and is my own work. I understand that I should not submit work that has not been written by myself.</td>
</tr>
<tr>
<td>5.</td>
<td>I give permission for my data to be stored and processed in accordance with the GDPR (2018)</td>
</tr>
<tr>
<td>6.</td>
<td>I agree to my anonymised data being used in study specific reports and subsequent articles that will appear in academic journals as part of this study.</td>
</tr>
<tr>
<td>7.</td>
<td>I agree to take part in the above study.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of participant</th>
<th>Date</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name of researcher taking consent</th>
<th>Date</th>
<th>Signature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Research into Academic Misconduct – Analysis of Student Work

Dear student

Do you want to take part in some research that is tackling some of the challenges relating to plagiarism and contract cheating? Would you be willing to submit an assignment for analysis?

If you’ve got assignments that have already been marked, and you know they are genuinely your own work (not copied from elsewhere or written by someone else), please get in touch with me, Clare Johnson. I am developing software that checks student submissions for flags that may indicate plagiarism, and to do this fairly I need a selection of genuine student work to establish baselines. There is no risk to your existing grades, and all work will be anonymised and aggregated for reporting purposes. You should not submit any work if you know that it is not genuinely your own.

If you would like to help with this important work, please send me an email to clare.johnson@southwales.ac.uk and I will forward to you the participant information sheet with further details, along with a consent form. You will then need to send me your assignment in Word format, via email and I will add this to my samples.

If you have any questions about the research, please feel free to contact me. I hope to hear from you soon.

My name is Clare Johnson and I am currently doing research into plagiarism and contract cheating. I am interested in finding out more about students’ perceptions of academic integrity and the causes of academic misconduct. My background is in cyber security and through my work with digital forensics I have developed a software tool that looks at the authorship of student assignments. I hope this tool will help assessors identify work that is authentic and that which is not.

In order to continue this important work, student assignments are needed for establishing baselines within the tool. If you are willing to submit an assignment for this purpose please contact clare.johnson@southwales.ac.uk. All work will be anonymised prior to analysis, and any results will be aggregated. Thank you.

If you have any questions about the research, please feel free to contact me. I hope to hear from you soon.
Appendix H: ‘Clarify’ code

The programming code for ‘Clarify’ is stored on a private GitHub repository which was made available for examiners on request. Information on accessing this code has been redacted.

A screen shot of the index page is included here and a sample of the output from the case study files can be found in Appendix E.
Appendix I: Published works

This appendix lists the following published works in order of publication:


Copies of the articles have been redacted for publication.