Biometrically Linking Document Leakage to the Individuals Responsible

Abdulrahman Alruban[1, 2] (✉), Nathan Clarke[1, 3], Fudong Li[1, 4] and Steven Furnell[1, 3, 5]

[1] Centre for Security, Communications and Network Research, Plymouth University
Plymouth, UK
[2] Computer Sciences and Information Technology College, Majmaah University, 11952
Saudi Arabia
[3] Security Research Institute, Edith Cowan University
Perth, Western Australia
[4] School of Computing, University of Portsmouth
Portsmouth, UK
[5] Centre for Research in Information and Cyber Security, Nelson Mandela University
Port Elizabeth, South Africa
{abdulrahman.alruban, n.clarke, fudong.li,
steven.furnell}@plymouth.ac.uk

Abstract. Insider threats are a significant security issue. The last decade has witnessed countless instances of data loss and exposure in which data has become publicly available and easily accessible. Losing or disclosing sensitive data or confidential information may cause substantial financial and reputational damage to a company. Whilst more recent research has specifically focused on the insider misuse problem, it has tended to focus on the information itself – either through its protection or approaches to detect leakage. In contrast, this paper presents a proactive approach to the attribution of misuse via information leakage using biometrics and a locality-sensitive hashing scheme. The hash digest of the object (e.g. a document) is mapped with the given biometric information of the person who interacted with it and generates a digital imprint file that represents the correlation between the two parties. The proposed approach does not directly store or preserve any explicit biometric information nor document copy in a repository. It is only the established correlation (imprint) is kept for the purpose of reconstructing the mapped information once an incident occurred. Comprehensive experiments for the proposed approach have shown that it is highly possible to establish this correlation even when the original version has undergone significant file modification. In many scenarios, such as changing the file format r removing parts of the document, including words and sentences, it was possible to extract and reconstruct the correlated biometric information out of a modified document (e.g. 100 words were deleted) with an average success rate of 89.31%.

Keywords: Digital forensics; biometrics; insider misuse; data leakage.

1 Introduction

It is deeply worrying for organisations when data exposure originates from an authorised individual (e.g. an employee or contractor) who misuses their legitimate access, and the potential for adverse impacts, in this case, is typically higher than that of access by outsiders [1–3]. Insiders are more likely to bypass security controls while outsiders, who typically have limited knowledge of internal infrastructure in a given case, pose a significantly smaller threat. Identifying such criminals, especially if the digital forensics process leads to the presentation of findings in legal proceedings, is a challenging and crucial task. Therefore, one of the aims of the digital forensics process is to produce and test a hypothesis about who did what, where, when and how in relation to an incident under investigation.

Existing methods and tools used by investigators to conduct examinations of digital crime significantly help in collecting, analysing and presenting digital evidence. Essential to this process is investigators establishing a link between the notable/stolen digital object and to the identity of the individual who used it; as opposed to merely using an electronic record or a log that indicates the user interacted with the object in question (evidence). This is a challenging task because it is currently difficult for digital forensic investigators to prove, to the appropriate standard in a court of law, that a specific human used a digital object (e.g. a document or image) at a particular time. An underlying assumption is that the identified computer account—as an example, of which the misuse occurred belongs to the individual who perpetrated the attack. However, with generally poor password use (e.g. shared or stolen accounts) and specific malicious intent, this is unlikely to be true. Thus, correlating such a link is key to identifying the individual(s) responsible.

This paper presents an approach that transparently acquires biometric signals from individuals as they naturally interact with the system, and tries to correlate their biometric information with the objects that they interact with, such as documents, email messages and photographs. In this manner, the biometric information of the last individual to access a digital object will be linked to it. Subsequent misuse of such information, through disclosure, for example, would enable an organisation to process the digital object, recover the biometric identifiers and identify the last employee who accessed it.

The remainder of the paper is organised as follows: Section 2 highlights the related work in the area of action logs and watermarking. Section 3 introduces the proposed approach, including the core process. Section 4 presents the experimental analysis and evaluates the robustness of the proposed method. Section 5 discusses the findings and possible directions for future work, and section 6 provides concluding remarks.

2 Related Work

The current solution for detecting insider misuse involves a layering of security countermeasures that includes comprehensive logging of servers (including authentication requests) so that logs can be correlated to understand who was using what machine at

what time, resulting in specific actions on the network [4–8]. Assuming encryption is in place, proxy-based network decryption and storage of network traffic is required to identify the misuse (possibly over prolonged periods of time). If third-party encryption is used, it can be challenging to decrypt and perform a deep inspection of the captured traffic [9–11].

A limited number of studies have tried to leverage soft biometric signals to detect malicious insiders' activities [12, 13]. Both studies proposed systems that employ the use of human bio-signals such as electroencephalography and electrocardiogram to detect insiders' malicious activities. For detection, they measure the difference in bio-signals deviations between normal and malicious activity phases. Although both systems deployed their approaches in real-life scenarios and achieved high detection accuracy, the experimental setup relies on users wearing a headset that continuously monitors bio-signals and a finger sensor to capture them. However, it is both unrealistic and non-user-friendly to wear these sensors in real life continuously.

Other researchers have employed steganography and watermarking techniques to embed specific data that could point to the action generator [14–17]. While the nature of conventional watermarking or steganography processes is not to modify the digital object in a manner that is noticeable, it does nonetheless modify the document. There may be situations where this modification is not desirable, for instance when preserving the integrity of the object is crucial.

Therefore, the proposed approach in this study seeks to provide a mapping technique between the digital object and biometric identifiers, storing the mapped information alongside document identifiers in a centralised storage repository. When the mapped (imprinted) objects are recovered or analysed, the information stored in the repository is used to recover the biometric information, which is subsequently used to identify the user. The key advantage of this approach is that the underlying digital object is not modified in any way, in contrast to the aforementioned watermarking studies. Also, no explicit biometric information is stored as only the correlation that points to locations within the imprinted object are preserved.

3 The Proposed Approach

The proposed approach takes advantage of Locality Sensitive Hashing (LSH) schemes to generate a less sensitive representation to modification of the document (text). In general, LSH algorithms are mainly used for dimensionality reduction by mapping high dimensional input space into lower dimensional space. A key difference between LSH based algorithms compared to cryptographic schemes is that the former is less sensitive to small changes on the mapped input space. In contrast to hash-based cryptographic schemes, which are designed for ensuring data integrity by maximising its sensitivity to the input space. Both methods map the input stream into a fixed output called digest (hash values). This study leverages LSH property of maximising the probability of a collision for similar inputs. This achieved by directly mapping the biometric feature vector representation of an individual with the computed LSH digest of a given document, this generates a digital—what is called 'imprint'

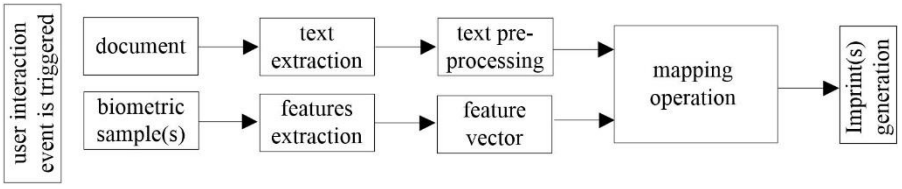

Fig. 1. Biometric information-document correlation generation pipeline.

file. The resulted imprint file represents locations within the computed LSH hash value, each of which corresponds to a respective portion of the digital biometric feature vector. The user's biometric samples from which the feature vector is computed (e.g. facial features, iris, keystroke analysis or behavioural profiling) are transparently and continuously captured – using suitable sensors – while the person is interacting with the computer. Finally, these generated imprints are stored in a centralised, secure database for later analysis when required. Fig. 1 illustrates the process of generating those imprint files which establish the correlation between the acquired biometric information of the corresponding person and the triggered document. Data leakage in the form of a document (whether posted on a public website or captured by the network) can be then analysed by processing the imprint file with the given 'leaked' document, which was already imprinted at some point before it was leaked, to reconstruct the mapped biometric feature vector. Once the sample is extracted, it can be processed by a biometric system in order to determine the last user who interacted with the object.

To illustrate how mapping the biometric feature vector with LSH digest works, assume that the following feature vector needs to be mapped with the given LSH digest as shown in Fig. 2.

	Value
Feature vector sample	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9]
LSH digest sample	[F1751BD78C133A4A9303D6365E78E4933D843436A7921120789B58138AFB927BF7DE]
Index	0……10……..20……..30……..40……..50……..60……

Fig. 2. Examples of a feature vector and TLSH digest sample

In this example, each value (digit) of the feature vector exists in more than one location within the hash digest. The sample digest in this figure was computed using TLSH scheme which outputs 70 hexadecimal characters long (35 bytes). TLSH is a type of LSH schemes developed by TrendMicro [18]. In mapping, "0" is located in two locations; 18 and 47. In the same manner, the mapping process finds all matching locations for the remaining values of the given feature vector as shown in Fig. 3.

F.V.	Matched index location within TLSH digest									
0	18	47								
1	1	4	10	44	45	54				
2	43	46	61							
3	11	12	17	19	22	31	32	36	38	55
4	14	29	35	37						
5	3	24	52							
6	21	23	39							
7	2	7	26	41	48	62	65			
8	8	27	34	49	56					
9	16	30	42	50	60					
	1st | 2nd imprints									

Fig. 3. Feature vector—LSH digest mapping matrix

By combining those mapped locations (one location from each row), this forms a single imprint. Hence, the total unique imprints that can be generated from the mapped indexes are two as highlighted in light green in Fig. 3. Therefore, using any of these imprints, it is possible to reconstruct the original (mapped) feature vector from the document by reversing the mapping process. The next subsection describes the correlation generation pipeline including the mapping process step.

3.1 Correlation Generation Pipeline

The generation process of the imprint file which associates individual's biometric signal with a document of interest involves six main steps starting with acquiring document's text and ending with generating the target imprint file.

Extracting document text

The document text is extracted from the file, and the text itself is processed, not the document file type. This approach makes it possible to imprint any document type so that its text can be extracted. For example, PDF, DOCX, TXT, HTML or even email messages can all be analysed, and their content can be parsed. Furthermore, the extraction process eliminates any text formatting; therefore, the subsequent steps of the imprinting process rely purely on the text.

Pre-processing the extracted text

In this phase, all extra spaces between words, lines, paragraphs and pages that exist in the text are removed and replaced with a single space. This ensures that the computed LSH digest is based only on the plain text, which means that if the document is maliciously manipulated later, for instance by adding extra spaces or page breaks, it will have low or even no effect on the computed hash value.

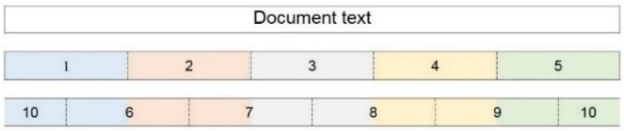

Fig. 4. Slicing document's text into 10-overlapped-folds.

Computing the LSH value of the text

The LSH value can be computed by using one of the known open-source algorithms, including Ssdeep, Sdhash, Nilsimsa or TLSH [19–21, 18]. It is well established that TLSH is more robust than the other schemes regarding the digest entropy, collision likelihood, as well as against manipulation attacks (e.g. removing, swapping, and inserting words) [22]. Therefore, the TLSH algorithm was chosen for use in this study to compute the hash digest of the extracted text. Also, two approaches can be used to compute the hash digest of the document as follows:

a) Only a single hash digest is computed for the whole document, this makes the imprinting process much faster and stores fewer data in the database as only one digest is used to generate the correlation with the biometric signal.

b) Hashing the text using a different resolution to produce multiple digests per document, for example, per page, half page, and a paragraph or using k-overlapped-folds of the examined document as illustrated in Fig. 4. It presents how document text is sliced into 10-overlapped-folds each of which is processed separately and its LSH value is computed.

In this study, methods (a) and (b) are both examined and evaluated against different possible attack vectors as detailed in Section 4.

Also, another LSH hash digest is computed (using, for instance, Nilsimsa) and stored in a centralised database to be used later to locate the associated imprint file when a questioned document is queried. Besides, the biometric signal is hashed using Secure Hash Algorithm (SHA) digest and stored as well. SHA is used for checking the integrity of extracted biometric signal. The reason for using another LSH algorithm is to avoid storing the same LSH digest which was used for generating the imprint. This ensures that having only the imprint in the database without the correlated document makes it impossible to reconstruct the related biometric information.

Mapping feature vector with Hash digest value

The feature vector and the LSH hash value of the text are mapped to its equivalent location in the text LSH hash value to retrieve the possible locations where they match as described previously in this section.

Generating the imprints

By retrieving the locations of each character of the feature vector with the object, it becomes possible to generate the imprints based on the obtained list of indexes, which means that multi-imprints of the whole feature vector can be generated by combining those positions.

3.2 Recovery algorithm

The recovery algorithm to extract and reconstruct the imprinted biometric information out of a questioned document in the case of information leakage–shares the same steps 1-3 of the imprinting process that listed above. This followed by the following steps:

a) The questioned document hash digest is computed (e.g. Nilsimsa) as input to the next step.
b) The related-stored imprint file is retrieved by querying the centralised database–using the computed hash digest–where previously generated fingerprints and imprints for all documents are stored.
c) The retrieved imprint file is mapped with the computed LSH value of the document in question, and the correlated biometric signal is reconstructed out of those mapped locations.
d) To validate the integrity of the reconstructed biometric signal, its SHA digest is compared against the stored digest generated when the imprint was created.

After explaining how the imprinting and retrieving techniques of the proposed approach work, the next section investigates the feasibility of imprinting biometric information with documents and later recovering them (even after the text is modified).

4 Experimental Analysis

The fundamental research question concerning the imprinting of the biometric signature is how robust the approach is, given subsequent modification of the document – arguably the key attack vector against this approach. An insider who intends to leak a confidential document could maliciously manipulate its content in order to destroy any tracks to avoid being traced. Therefore, to examine the feasibility and effectiveness of the proposed approach, real leaked documents from WikiLeaks were chosen for experimental purposes. WikiLeaks is an international non-profit organisation that publishes secret information, news leaks and classified media provided by anonymous sources [23]. In 2009, it released more than six thousand reports commissioned by the United States Congress. These reports are classified as confidential documents and are now publicly available and accessible online in the form of text files [24]. Table 1 provides statistical information about the used dataset. Leaking repositories such as WikiLeaks and The Intercept typically perform some kind of modifications to the leaked documents. For instance, they watermark uploaded documents and files with extra information such as document ID, date, website address or logo [25].

Table 1. Corpus statistics

File size distribution(KB)	#of docs	Doc content	Min	Max	Average
1-99	4,920	Chars.	1,288	874,548	47,345
100-199	853	Words	233	155,614	8,873
200+	227	Lines	38	16,160	981
Total	6,000	Pages	1	622	34

A number of experiments were designed and conducted to evaluate the proposed approach in such scenarios that consider malicious intent with regard to any possible modification could be performed on the document. The first experiment maps the biometric feature vector with the computed text TLSH digest and retrieves it. The goal is to compute the possible number of imprints that can be generated from the mapping process. In addition, a total of twenty-one attacks were developed. This includes, file, formatting and text-based manipulation methods. These attacks critically examine the effectiveness of possible modification attacks on the imprinted documents and inspect how such attacks could affect the retrieval performance of the mapped biometric information. These developed attacks are classified into three main categories: file-type conversion, formatting change and content manipulations, as listed in Table 2.

Table 2. Possible document manipulation methods

File-type conversion	Formatting change	Content manipulation
1. PDF to .docx	9. Font resizing	14. Deleting words
2. PDF to .txt	10. Font type changing	15. Deleting sentences
3. PDF to Image	11. Colour changing	16. Deleting lines
4. Docx to PDF	12. Text highlighting	17. Swapping words
5. Docx to txt	13. Line and para spaces	18. Swapping sentences
6. Txt to PDF		19. Swapping lines
7. Txt to .docx		20. Substituting synonyms
8. Txt to Image		21. Inserting new words

The used biometric feature vectors, in the imprinting process, represent real facial features. Fisherfaces feature extraction algorithm is used to compute these vectors for the captured users' faces images [26]. The dimensions of the generated feature vector when using Fisherfaces algorithm is small compared to deep learning approaches as the length of the vector is a prime factor when performing imprinting process. The resulted vector is 4-dimensions with the length of 60 digits. The chosen vector includes frequency of all digits (0, 1, 2, 3, 4, 5, 6, 7, 8, 9) as well as '-' sign, this to ensure that this study covers all possible numbers within the mapping process.

In all manipulation methods above, the original document TLSH value is computed before it is modified, and the resulting digest is then imprinted with the biometric information. After that, the manipulation methods are applied to the imprinted documents. Finally, the TLSH value of the modified version is computed again and compared to the original one. As long as the original text has not changed, the full mapped biometric feature vector should be successfully retrieved by reversing the imprinting process. However, this is not always the case, since a leaked document is highly likely to have been manipulated or modified. Consequently, the computed hash value is directly affected, to what degree is depending upon the scale of modification. Fortunately, TLSH is less sensitive to small changes than cryptographic hashing algorithms, such as SHA, since a small modification in the input drastically changes the output computed digest. This is the so-called avalanche effect.

6efa3f05f084127249ebe7e0b37ffdda41db9ceacfbb65c04cd7de6a
SHA256 digest of the original document
49ca48a970f02c40cf85667d1708416ccca84de06d65467856aa3ef1
SHA256 the digest of the modified document

77F1866D9E10AF925F4228F3475961F8C0DAB4751388000565A1B8571D67C7E1F5A6FE1BE78C133A4A9303D6
365E7CE8933D843437A7D21120789B58238AFB927BF7DE
TLSH digest of the original document

4DF1856D4E106F925F4224F7476961F8C0DBB0751388001565A178571D67C7E0F1AAFF1BE78C133A0A9303D6
365E68E5A33D843437A791152078 9B58238AFB927BF7EE
TLSH digest of the modified document

Fig. 5. Samples of computed document hash digests using SHA256 and TLSH

In contrast, all similar digest schemes have the property that a small change to the file being hashed results in small change to the hash [18]. For example, Fig. 5 shows two samples of computed document hash digests using SHA256 and TLSH. Each presents two values: one for the original document and one for the modified version of the same document. It is clearly shown that the digest of the modified document computed bySHA256 is entirely different to the originals. While the TLSH digest is only slightly affected, the red characters are those changed while the others remain the same with exact locations. Therefore, the TLSH can be used in our approach to give a less sensitive representation for the whole document.

Fig. 6 shows the averaged distribution of the number of imprints generated per document for the examined 6,000 documents in the dataset. The histogram indicates that most of the imprinted documents generated more than three imprints. The number of the obtainable imprints mainly depends on the generated TLSH digest entropy and digits frequency. The rate of the entropy and frequency differ from one document to another as this is natural property of hash schemes. Although multiple imprints per document were generated as the figure illustrates, in fact, only one imprint is needed to successfully reconstruct the biometric information. Indeed, having multiple imprints for a given document significantly increases the probability of recovering the mapped information even after the document is exposed to manipulation.

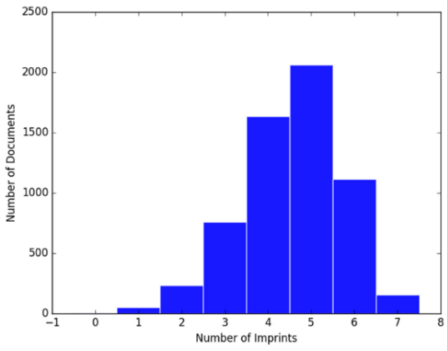

Fig. 6. Distribution of the generated imprints per document.

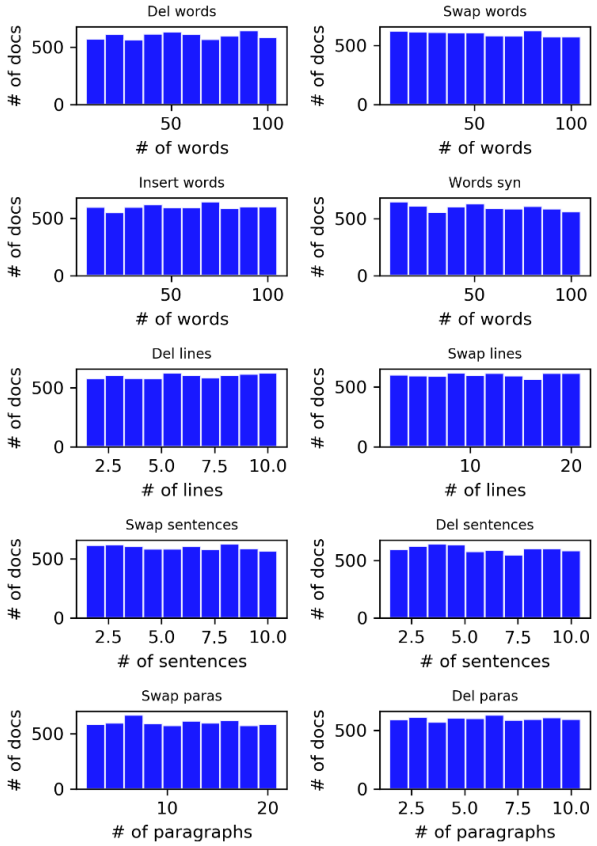

Fig. 7. Distribution of change rate among dataset documents.

The experimental results of the developed method indicate that the proposed approach is resistant and robust against both file-type conversion and formatting change attacks with an accuracy of 100%. Since the nature of these modification methods does not change the actual text or content which is fed into the LSH algorithm, therefore, the mapped biometric signal is fully retrievable even when the text format or file-type is changed, including converting the document into an image file type. However, in such a case, Optical Character Recognition (OCR) technologies could be used to analyse and convert the image content (printed text) into machine-encoded text. In this study, test documents were converted into images (JPEG) to simulate such an attack, and a Tesseract-OCR engine was used to read all those images and recognise and extract the embedded text [27]. As long as the OCR was able to recognise the correct text, which it did, the integrity of the text can be preserved compared to its original version.

For the content manipulation attacks, random settings were configured for the rate of modification, as Fig. 7 illustrates, ranging from 1 to 100 for word-type attacks and 1 to 20 for line and paragraph attacks. As this rate increases, the number of changes rises as well. For instance, in the case of the word-deleting attack, a number of ran-

dom words between (1, 100) are deleted from each document in the dataset. Also, this applies to all other attacks that fit in the same category.

Table 3 presents the results of retrieving the mapped feature vector under the content manipulation attack methods. In addition, TLSH uses a distance score of '0', which indicates that the files are identical (or nearly identical), while scores above that represent a greater distance between the examined documents. A higher score should represent that there are more differences between the documents [18].

From the data in Table 3, it can be seen that given the capability of recovering biometric identifiers under significant levels of modification—such as deleting 100 words—it is still possible to regenerate the established correlation between the biometric information and the imprinted document with a success rate of (89.31%). In addition, Fig. 8 illustrates how the accuracy changes along with a defined number of deleted words. Two levels of hashing resolutions were applied on the examined documents, one hash digest per document and multi-hash digest using 10-overlapped-folds per document. The overall accuracy is improved when multi-hash digests are generated. In general, a document is counted as correctly identified (feature vector is retrieved) if at least one imprint is perfectly extracted from the imprinted feature vector even when the computed hash digest is not identical to the one from which the original correlation where established.

Table 3. Content manipulation attack methods experimental results.

No	Attack type	Rate (number)	#of retrieved F.V.	Score (%)	TLSH diff (original/modified)[1]		
					Min	Max	Avg.
1	Del words	1-100	5,359	89.31	0	217	8
2	Swap words	1-100	5,464	91.06	0	82	7
3	Insert words	1-100	5,304	88.40	1	471	33
4	Words syn.	1-100	5,751	95.85	1	465	30
5	Del lines	1-10	2,708	45.13	7	466	43
6	Swap lines	1-10	2,637	43.95	7	874	71
7	Swap sentences	1-10	5,929	98.81	0	30	3
8	Swap paras	1-10	2,853	47.55	5	125	26
9	Del paras	1-10	2,767	46.11	5	149	26
10	Del sentences	1-10	4,915	82.00	1	788	15
11	Multi attacks[2]	1-10	3,828	64.00	1	456	31

[1] TLSH diff is distance score between two digests (texts)

[2] A number of attack methods are randomly chosen

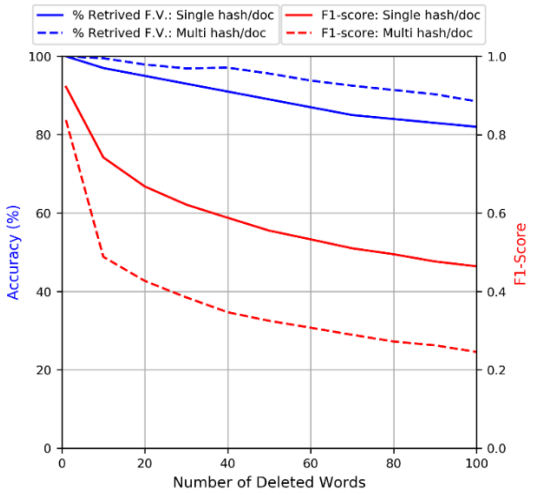

Fig. 8. Averaged accuracy and F1-Score for the deleted words attack.

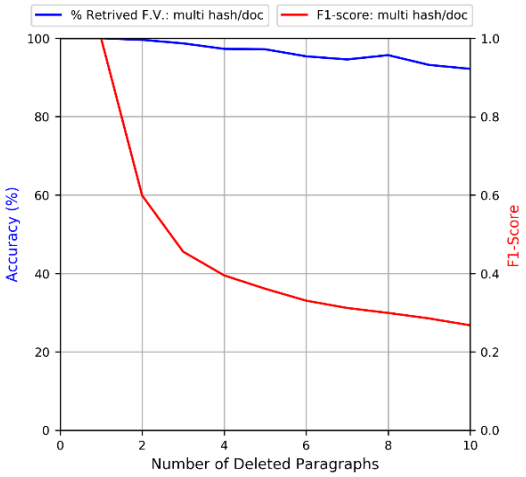

Fig. 9. Averaged accuracy and F1-Score for the deleted paragraphs attack.

Furthermore, paragraph attacks (swap and delete methods) have scored low rates, with 47.55% and 46.11% respectively. Indeed, removing a number of paragraphs from the document significantly affects the computed hash digest to a greater degree than other types of modification, such as deleting words or sentences. This can be improved by changing the hashing resolution (i.e. using k-folds). For instance, instead of hashing the whole document and generating a single hash digest, multiple digests are computed for the document, for example per page, half page or paragraph, and correlating the biometric information with the resulted hashes. Fig. 9 shows the aver-

aged accuracy and F1-score for the deleted paragraphs attack using 10-overlapped-folds. The overall accuracy is higher than the single hash digest per document approach as it scored 93%. In contrast, the achieved F-score is not high as it computed for all the generated imprints, while only one valid retrieved imprint of a given document is needed to reconstruct the mapped biometric information. Moreover, chances for recovering the correlated biometric signals vary based on the type and scale of attack vector. However, in many leakage cases, the leaked document might not be exposed to a severe modification. Hence, reconstructing the biometric sample is highly likely to be possible and, as a result, the source of leakage can be identified.

5 Discussion

The most obvious finding to emerge from this study is that the underlying digital objects, documents in this case, are not modified in any form. In addition, the proposed approach also disassociates any biometric information from the digital object itself, thereby minimising any attacks on the biometric data. Which means that the biometric single is not stored by any means in a database, only its correlation to the imprinted object (document/text in this case) is preserved in the imprint file. Thus, it becomes useless without having the imprinted document in presence for the recovery process, since the imprint file that correlates the object with the related biometric signal only contains those locations within the document where the signal can be extracted from. Besides, it allows for larger volumes of information to be imprinted, making it more suitable for digital objects when greater levels of information need to be correlated (i.e. multimodal biometric samples). It does, however, introduce the need for a centralised repository which will grow as users interact with objects and thus requires configuration and management.

Although the above investigation has critically examined the proposed approach against possible malicious attacks and showed robustness and strength, a number of challenges exist and require further research. These include the ability to automate the process of capturing the biometric signal and detecting user interaction with the object instantly, along with establishing the correlation with the interacted object. This requires the development of a smart and active agent that continually captures an individual's biometric information (using a camera in the case of facial information) and performs the imprinting process. Furthermore, the proposed approach raises important privacy concerns for those individuals who are monitored by the system. In which processing, transmitting and storing the biometric samples into a centralised database require a high level of confidentiality and sufficient resources. This obviously needs to be investigated in depth in the future work. More broadly, research is also needed to determine the ability to utilise a broader range of digital objects. Differing objects have varying degrees of stability due to their structure. For example, executable files and their underlying data structure can change considerably given small alterations to a file, in contrast to text. Therefore, the proposed approach needs to be examined for such file-types to fully measure its usefulness and robustness. Also, further study needs to be carried out regarding the ability to utilise soft biometric features such as the gender, age and even race of individuals to increase the discriminative ability and provide more reliable information to the investigator.

6 Conclusion

This paper has introduced a proactive approach to aiding an incident investigator to establish and examine a case of insider misuse, particularly with respect to information leakage, and could increase the likelihood of the evidence being admissible in a court of law. This study has shown that it is possible to successfully recover biometric information even under significant modification attacks. Rather than requiring the complete digital object, it is possible to recover the necessary information with even a modified version of the questioned document.

7 Acknowledgements

This research was undertaken with the support of the Majmaah University, Majmaah city, Saudi Arabia.

References

1. Titcomb, J., WikiLeaks releases thousands of hacked Macron campaign emails, 2017. [Online]. Available: http://www.telegraph.co.uk/news/2017/07/31/wikileaks-releases-thousands-hacked-macron-campaign-emails/. [Accessed: 07-Sep-2017].
2. WikiLeaks publishes 'biggest ever leak of secret CIA documents,' 2017. [Online]. Available: https://www.theguardian.com/media/2017/mar/07/wikileaks-publishes-biggest-ever-leak-of-secret-cia-documents-hacking-surveillance. [Accessed: 09-Sep-2017].
3. Moshinsky, B., LEAKED DOCUMENT: Bank of England has 'significant concern' over post-Brexit approval for Deutsche Bank's UK branch, 2017. [Online]. Available: http://uk.businessinsider.com/bank-of-england-document-deutsche-bank-post-brexit-uk-2017-8. [Accessed: 07-Sep-2017].
4. Rahayu Selamat, S., S. Sahib, N. Hafeizah, R. Yusof, and M. Faizal Abdollah, A Forensic Traceability Index in Digital Forensic Investigation, J. Inf. Secur., vol. 4, no. 1, pp. 19–32, 2013.
5. Homem, I., S. Dosis, and O. Popov, LEIA: The Live Evidence Information Aggregator: Towards efficient cyber-law enforcement, in *World Congress on Internet Security (WorldCIS-2013)*, 2013, pp. 156–161.
6. Magklaras, G., S. Furnell, and M. Papadaki, LUARM – An Audit Engine for Insider Misuse Detection, Int. J. Digit. Crime Forensics, vol. 3, no. 3, pp. 37–49, Jan. 2011.
7. Homem, I., S. Dosis, and O. Popov, The Network Factor in Proactive Digital Evidence Acquisition, Int. J. Intell. Comput. Res., vol. 6, no. 1, pp. 517–526, 2015.
8. Quick, D. and K.-K. R. Choo, Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata?, Digit. Investig., vol. 10, no. 3, pp. 266–277, Oct. 2013.
9. Pilli, E. S., R. C. Joshi, and R. Niyogi, Network forensic frameworks: Survey and research challenges, Digit. Investig., vol. 7, no. 1–2, pp. 14–27, Oct. 2010.
10. Khan, S., A. Gani, A. W. A. Wahab, M. Shiraz, and I. Ahmad, Network forensics: Review,

taxonomy, and open challenges, J. Netw. Comput. Appl., vol. 66, pp. 214–235, May 2016.
11. Birk, D. and C. Wegener, Technical Issues of Forensic Investigations in Cloud Computing Environments, in *2011 Sixth IEEE International Workshop on Systematic Approaches to Digital Forensic Engineering*, 2011, pp. 1–10.
12. Hashem, Y., H. Takabi, M. GhasemiGol, and R. Dantu, Towards Insider Threat Detection Using Psychophysiological Signals, in *Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats - MIST '15*, 2015, vol. 6, no. 1, pp. 71–74.
13. Almehmadi, A. and K. El-Khatib, On the Possibility of Insider Threat Detection Using Physiological Signal Monitoring, in *Proceedings of the 7th International Conference on Security of Information and Networks - SIN '14*, 2014, pp. 223–230.
14. Bouslimi, D. and G. Coatrieux, A crypto-watermarking system for ensuring reliability control and traceability of medical images, Signal Process. Image Commun., vol. 47, pp. 160–169, Sep. 2016.
15. Chaabane, F., M. Charfeddine, and C. Ben Amar, A survey on digital tracing traitors schemes, in *2013 9th International Conference on Information Assurance and Security (IAS)*, 2013, pp. 85–90.
16. Macq, B., P. R. Alface, and M. Montanola, Applicability of watermarking for intellectual property rights protection in a 3D printing scenario, in *Proceedings of the 20th International Conference on 3D Web Technology - Web3D '15*, 2015, pp. 89–95.
17. Alruban, A., N. Clarke, F. Li, and S. Furnell, Insider Misuse Attribution using Biometrics, in *Proceedings of the 12th International Conference on Availability, Reliability and Security - ARES '17*, 2017, pp. 1–7.
18. Oliver, J., C. Cheng, and Y. Chen, TLSH -- A Locality Sensitive Hash, in *2013 Fourth Cybercrime and Trustworthy Computing Workshop*, 2013, no. November 2013, pp. 7–13.
19. Kornblum, J., Identifying almost identical files using context triggered piecewise hashing, Digit. Investig., vol. 3, no. SUPPL., pp. 91–97, Sep. 2006.
20. Roussev, V., Data Fingerprinting with Similarity Digests, in *IFIP Advances in Information and Communication Technology*, vol. 337 AICT, 2010, pp. 207–226.
21. Damiani, E., S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, An Open Digest-based Technique for Spam Detection, Proc. 2004 Int. Work. Secur. Parallel Distrib. Syst., vol. 1, no. 1, pp. 559–564, 2004.
22. Oliver, J., S. Forman, and C. Cheng, Using Randomization to Attack Similarity Digests, 2014, pp. 199–210.
23. WikiLeaks. [Online]. Available: https://wikileaks.org. [Accessed: 05-Sep-2017].
24. A billion in secret Congressional reports, 2009. [Online]. Available: https://wikileaks.org/wiki/Change_you_can_download:_a_billion_in_secret_Congressional_reports. [Accessed: 04-Sep-2017].
25. The Intercept. [Online]. Available: https://theintercept.com/. [Accessed: 05-Sep-2017].
26. Belhumeur, P. N., J. P. Hespanha, and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell., vol. 19, no. 7, pp. 711–720, 1997.
27. Smith, R., An Overview of the Tesseract OCR Engine, in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 2007, pp. 629–633.