Title	**Protecting digital data privacy in Computer Forensic Examination**
Author(s)	**Law, FYW; Chan, PF; Yiu, SM; Chow, KP; Kwan, MYK; Tse, HKS**
Citation	**The 6th International Workshop on Systematic Approaches to Digital Forensic Engineeringi in conjunction with IEEE Security and Privacy Symposium (IEEE/SADFE 2011), Oakland, CA., 26 May 2011.**
Issued Date	**2011**
URL	**http://hdl.handle.net/10722/139988**
Rights	

Protecting Digital Data Privacy in Computer Forensic Examination

Frank Y.W. Law, Patrick P.F. Chan, S.M. Yiu, K.P. Chow, Michael Y.K. Kwan, Hayson K.S. Tse, Pierre K.Y. Lai

The University of Hong Kong, Hong Kong

***Abstract*—Privacy is a fundamental human right defined in the Universal Declaration of Human Rights. To enable the protection of data privacy, personal data that are not related to the investigation subject should be excluded during computer forensic examination. In the physical world, protection of privacy is controlled and regulated in most countries by laws. Legislation for handling private data has been established in various jurisdictions. In the modern world, the massive use of computers generates a huge amount of private data and there is correspondingly an increased expectation to recognize and respect human rights in digital investigation. However, there does not exist a forensically sound model for protecting private data in the context of digital investigation, and it poses a threat to privacy if the investigation involves the processing of such kind of data. In this paper, we try to address this important issue and present a cryptographic model designed to be incorporated into the current digital investigation framework, thereby adding a possible way to protect data privacy in digital investigation.**

***Index Terms*— Computer Forensics, Data Privacy, Data Protection, Digital Investigation**

I. Introduction

Digital data privacy is the relationship between technology and the legal right to, or public expectation of privacy in the collection and sharing of data [2]. With the advancement in information technology, data privacy is no longer limited to paper information and has extended to various kinds of electronically stored information such as emails, faxes, instant messages, electronic word documents, voice mails, digital images, spreadsheets etc. These kinds of data have increasingly become the focus of investigation in both criminal and civil cases. Digital investigators unavoidably require access to private information in the context of examining heterogeneous digital storage devices. This conventional investigation and the access to private data is normally justified on the occasion that the device belongs to a single user, who is the implied owner of the data. However, in some situations where the computing environments are shared or digital storage devices are sharing amongst multiple users, the investigation becomes more complicated and the data owner may not want to release their information for investigation. As a consequence, it is sometimes impractical to acquire a full image without the consent of the data owner. Notwithstanding, it is recognized that partial or selective file copying may be considered as an alternative in such circumstance in accordance with some practical guidelines [3].

The usual computer forensic examination process emphasizes the importance of data integrity and requires the acquisition process to obtain a full bit-stream image of the original storage media content. This approach preserves all the data of the target device without considering the issue of data privacy. Criminal laws often empower the investigator to examine the information to determine its relevance to the case and therefore justify it seizure. Regardless, when considering civil search or seizure orders, the full collection of data may expose private information that is not related to the investigation subject. For example, there are instances when investigation is required on a corporate email server. The email server only contains a small portion of emails that are related to the investigation subject but the process of obtaining a full image of the server allows the digital investigator to access millions of emails which are not related to the case. Regardless of the investigator focusing on email messages that contained the subject's email address or defined keyword hits, there may be occasions during which other messages are read by the investigator. Furthermore, a data breach can occur in myriad ways and may lead to violations of privacy. Counter-forensic privacy tools which locate and destroy private records are therefore developed to address the concern about recovering sensitive private data from computer systems [5]. It is observed that data privacy protection is required to be addressed in the context of computer forensics examination and a systematic approach is needed to assist the digital investigator in handling private digital data appropriately and effectively.

Unfortunately, the majority of existing published digital forensics investigation models or procedures have not incorporated the procedure for supporting data privacy protection. For instance, in the DFRWS framework [4], digital investigation covers the Identification, Preservation, Collection, Examination, Analysis and Presentation of digital

Manuscript received February, 2011. Corresponding author: Frank Law (e-mail: ywlaw@cs.hku.hk).

evidence. It focuses on the technical aspects in collecting, examining and explaining the hypothesis of incidents without incorporating the matter of data privacy. Therefore, without the inclusion of data privacy protection into the existing digital forensics investigation model, private information can only be protected through individual operating procedures which limit the search for evidence to the goal of investigation.

In this paper, we address the proper protection of private data during digital investigation, in the hope of providing a set of forensically sound procedures and proposing a practical and efficient approach to the task. The rest of the paper is organized as follows:

Section 2 highlights the difficulties in handling private digital data when compared to the physical world. Current practices on private digital data protection in the context of digital investigation are presented in Section 3. Then we propose a possible solution in Section 4. Section 5 describes a case study based on a simulated police investigation. Section 6 concludes the paper.

II. Data Privacy Protection at investigation

In the physical world, people are increasingly concerned about the handling of their personal data in the context of investigation because any disclosure may significantly breach their privacy. Under common data protection principles, private data could only be accessed when it matches certain pre-requisite criteria [6]. Nonetheless, exemptions are often granted to law enforcement officers for the purpose of prevention and detection of crime. In other words, data that are related to criminal acts are not protected. Though investigation normally focuses on the collection of information that is related to a specific crime, there may be a chance that the investigator will come across private information which is not related to the case.

Prior to the revolution of technology, securing privacy data was much easier because infringement of data privacy required much effort in extracting information from hard copies. Similarly, controlling and tracking how information was accessed by the investigator was easier because the scope and method of access were confined. The sharing of information was laborious and storage space was required if immense amount of materials were involved.

When it comes to the digital world, thousands upon thousands of digital files may be stored in a single digital storage medium. This greatly increases the potential points of information disclosure and there are instances in which private data were disclosed upon a loss of physical digital storage media, e.g. USB devices [23]. Unlike investigation in the physical world, digital investigators could access these kinds of private data sources in an effortless manner. With the advent of computer forensic tools, it is also simple to search and locate specific data sets, such as emails, credit card numbers, passport numbers, telephone numbers, identity card numbers, photographs, videos etc. in the context of investigation. In order to protect private data in the digital format, one method is by encryption, which helps prevent unauthorized disclosure of information. However, the difficulty lies on how to perform encryption on the data so that only relevant data can be retrieved by the investigator while other irrelevant data is not accessible.

Indeed, privacy protection requirements have an increasing impact in the real world. A number of privacy policies and legislations require the adoption of privacy protection measures when sensitive information is stored or processed. It is important to design solutions in response to this demand.

III. Current practices on digital data privacy

Kerr [8] opined that search and seizure of digital evidence should follow the Fourth Amendment of the US Constitution on discovery, seizure and searching. He commented that digital investigation processes should follow the procedures stipulated in the physical world investigation. As a result, the owner of a computer which was seized and searched should have a “Reasonable expectation of Privacy” as stated in Katz v. United States [17]. Indeed, there were laws [6,18,19] established to ensure that no greater invasion of privacy was permitted than under specific defined circumstances.

Nowadays, more and more people are becoming aware of their data privacy in the digital format. Many people started developing practices to remove sensitive digital data that are no longer used. The simplest method is to remove the data by deletion. However, there is much research to show that if digital data is deleted by the user, it does not mean that the data is securely removed from the storage media [9,10,11]. P. Stahlberg et al [7] identified threats to data privacy during computer forensic examination of database systems. The research revealed that users had little control over the persistence of deleted data in database systems and proposed specific techniques for secure record deletion and log expunction, making it more resistant to forensic analysis.

Indeed, digital traces may be permanently removed in the process of secure deletion. To prevent the recovery of deleted data, various documents or policies were published which detail the procedures for clearing, purging, or destroying digital data from its storage media [12,13,14,15]. One of the prevalent standards for secure deletion is the United States Department of Defense 5220.22-M standard for overwriting the data previously stored on magnetic storage media with a predefined set of meaningless data [16]. Apart from secure deletion, the method of encryption could also be employed to protect data in the digital format. Boneh et al [24] first adopted this method to remove data from files and backup logs. People who are concerned about the existence

of their private data on a computer could easily remove remnants on digital device by using various tools developed according to the aforementioned policies. The effectiveness of these types of tools has been evaluated by Geiger et al [22].

IV. Current practices on digital data privacy

In order to comply with common principles of computer forensics, it is observed that the entire digital investigation process should remain unchanged. The proposed approach should be the one adaptable to the current procedures and easily performed by the digital investigator. The core components of computer forensic investigation include obtaining a bit-stream image from the digital storage device, authenticating the evidence and analyzing the image to extract relevant digital evidence for reporting purpose. In the context of examination, the digital investigator could access the data content via computer forensic tools and reveal any relevant traces for proving a case. One obvious problem is that the image would replicate all the digital data, including the deleted data that exists on the storage media. With the assistance of standard computer forensic tools, one could easily inspect all data including logical files, deleted files or fragmented file data that exist within the image. With the assistance of the data owner, it may be possible to separate private data from the cloned image. However, it may take a very long time to view and segregate such information from the enormous amounts of digital data, and the identification of data may be error-prone under such an environment. For better accuracy and efficiency, the process that requires face-to-face interaction should be kept to a minimum.

Since deletion is not always feasible for computer forensic examination, another effective means to protect sensitive digital data is by encryption. Encryption is a straightforward and useful tool to protect the confidentiality of data. Recognizing the effectiveness of encryption, some vendors have employed encryption to protect important data [30,31]. However, it may create implications in the stage of analyzing data during computer forensic analysis because it is difficult to conduct searches on encrypted data. Though it may be possible for the examiner to obtain the encryption key for the purpose of data searching, all the data would be decrypted during the analysis stage, creating the risk of the examiner accessing private data.

In the proposed approach, it is suggested to apply encryption at both the data collection and data analysis stages to prevent the access of irrelevant data by the digital investigator at anytime during the examination. The core technology behind this method is related to the branch of research that allows one to search encrypted data using authorized keywords. To facilitate the searching of evidential data, index files will be built against all the content of the digital storage media. The approach is straightforward and involves the implementation of three modules. It first prepares a set of index files that allow digital investigators to search for relevant evidence about the case. The index files are basically some keywords that correlate with the digital data stored in the bit stream image. To prevent leakage of information from this point, the index files are encrypted by an encryption key provided by the data owner. During the examination, the investigator will be given a key constructed by the data owner. The key provided by the data owner will embed the keywords that are searchable by the investigator. Data not containing these keywords will not be accessible to the investigator. The index files will then be searched by the examiner through the keyword searching method. The sectors of data with relevant keyword hits will be extracted from the image for further examination by the investigator.

This method complies with the traditional bit stream acquisition method, but derives a way to properly handle large amounts of digital information in a forensically sound manner. The keyword approach also allows investigators to search for relevant information that is related to a specific issue. Unlike in the traditional method, this approach is specific and prevents access by the investigator to other digital data that are not relevant to the case.

Taking all these steps into account, it is proposed the following procedure for handling private digital data:

1. When investigators come across digital storage media where digital data privacy is one of the concerns, the content of the digital storage media should not be examined but a bit-stream image obtained via normal computer forensic processes.
2. The data owner generates an encryption key.
3. The image will then be scanned to build index files correlating the data content and its sector locations in the image. Encryption will then be applied on the index files to prevent possible leakage of information.
4. Before examining the data of the index files, the investigator will prepare a list of keywords which are relevant to the investigation. These keywords will be used to search for digital evidence from the acquired image.
5. The investigator obtains the searching key from the data owner and can search based on the set of authorized keywords. The image sector(s) where relevant keyword hits are recorded would be obtained. Related digital data could then be extracted out from the image for further examination.

V. Analysis of existing encryption schemes

One of the critical factors in this proposed approach is the speed of encryption during the data acquisition process. To identify a swift and secure encryption method, a number of existing schemes were compared. In [25] Boneh et al. provided three constructions for public-key searchable encryption. The first one is an efficient system based on a

variant of the Decision Diffie-Hellman assumption. The second system is less efficient using elements modulo a composite. The third system is based on general trapdoor permutations. Apart from the third system which is secure in the standard model, the other two are secure in the random oracle model. All three schemes are based on some identity-based encryption schemes. The three schemes are not employed in this paper as public-key encryption is not necessary in the context and the pairing operations involved in these schemes consume a considerable amount of CPU time.

In [26] Golle et al. defined a model for a conjunctive keyword search over encrypted data and presented the first scheme that conducts such searches securely. The security of their scheme is based on Decisional Diffie-Hellman assumption. Park et al. later extended this work to a public key encryption system [29]. Their constructions are based on decision bilinear Diffie-Hellman assumption, decision bilinear Diffie-Hellman inversion assumption and strong Diffie-Hellman assumption. The assumption of these two works is that the same keyword never appears in two different keyword fields and every keyword field is defined for every document. These two assumptions cannot be made in our context as the keywords are a set of unique words in the e-mail in our scenario. It is not reasonable to assume that all the e-mails have the same number of unique words.

Goh presented a secure indexing scheme in [28] which is based on bloom filter and pseudo-random function. A natural application of the secure index for searching encrypted data is provided in the paper. The secure index also supports advanced search queries.

The implementation of the proposed approach is based on the construction provided in Dawn et al [27]. Their scheme is provably secure and the construction is elegant. Only stream ciphers, HMAC and AES operations are involved in the scheme. Therefore, it was expected that it would be highly efficient and accurate. The implementation of system confirmed this expectation.

VI. The Implementation

To test the proposed approach, a system is implemented in Java following the concept of encrypted data searching [27]. The system comprises three modules and the system was tested against an image containing pure e-mail messages.

The first module is for building the index files. Before the investigator can search over the e-mails, the data owner, e.g. system administrator, needs to build an index file for each e-mail. The data owner needs to provide his keys, which will not be known by the investigator throughout the process, before the system builds the index files. The keys are generated beforehand using the system. The index files are encrypted and would not leak any information about the e-mails to the investigator.

The second module is the trapdoor calculation module. A trapdoor is essentially a piece of data that enables the investigator to search for a specific keyword over the encrypted index files. The trapdoor is provided by the data owner upon receiving a request from the investigator. The data owner needs to provide his keys again in order for the system to calculate the correct trapdoor. Therefore, only the data owner is able to generate the correct trapdoor. The investigator is not able to perform an arbitrary searching on his own initiative.

The third module is the searching module. After the investigator gets the trapdoor from the data owner, he is able to search over the encrypted index files using the trapdoor. The searching function can handle more than one keyword in a single query. The system will return the name of the index files that contain any of the specific keyword(s) represented by the trapdoor(s). The investigator can then ask the data owner to provide the corresponding e-mail files.

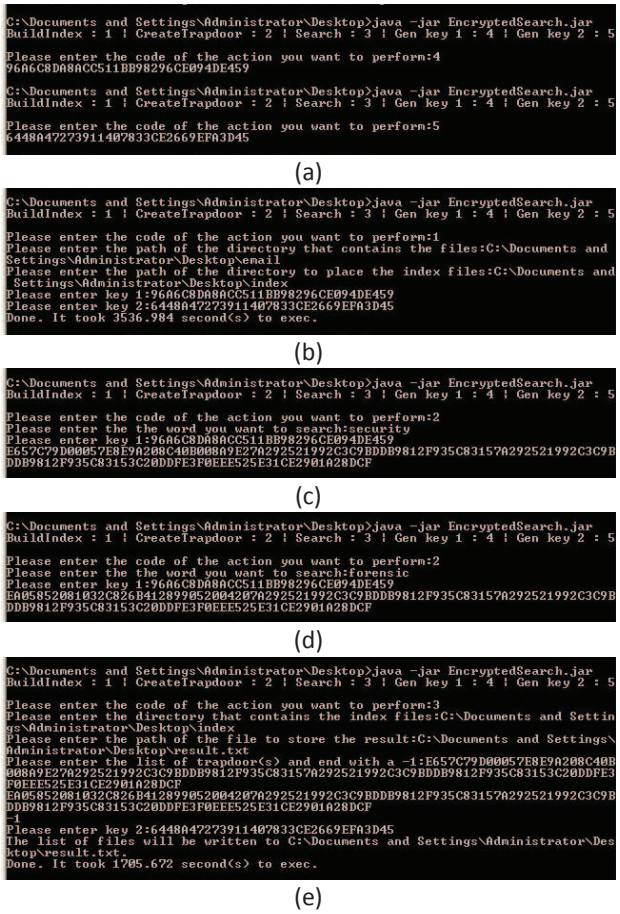

Figure 1. A sample running of the implemented system. (a) Generating the keys for encryption. (b) Building the indexes for searching. (c) Generating the trapdoor for the keyword security. (d) Generating the trapdoor for the keyword forensic. (e) Using the trapdoors and the keys, searching for the two keywords. The results are saved in a text file on the file system.

In order to analyze the performance of the system,

100,000 e-mails were prepared in which 25 percent of them contained the word "security", 25 percent of them contained the word “forensic” and 25 percent of them contained both “security” and “forensic”. Each e-mail comprises about 600 words. The system took 3536.984 seconds to build the index files. It took only 0.25 second to calculate the trapdoors for the words "security" and “forensic”. For the searching, it took 1705.672 seconds to search over all the e-mails and return the name of the e-mails that contained either “security” or “forensic”. Since the keywords appear nearly at the end of the e-mails as designed, this experiment is indeed testing the worst case scenario. The analysis was performed using a computer running Intel Core2Duo 2.66Ghz CPU with 2GB ram and 250GB hard disk. Figure 1 shows the screen captures of a demonstration.

The analysis showed that the system performed efficiently. More importantly, it takes little time to search over large amount of e-mails. The time it took to build the index is promising as well. The result shows that the system is ready to be used in practice due to its high efficiency and perfect accuracy.

VII. Case Analysis

To further evaluate the effectiveness of the proposed approach in practice, a simulated case was built to review its performance.

A corporate email server (120 Gb in size) containing suspicious emails related to a case of “Deception” was established. Apart from the suspicious emails, the email server stored an enormous amount of emails encompassing “sensitive” information that was totally unrelated to the case investigation. For example, some emails may have contained information that was subject to legal professional privilege, some emails may have contained sensitive trade information that may affect stock prices, some emails may have contained information relating to company plans, personnel movement, salary details, etc, which may affect the operation of the company.

With the amount of digital data involved and the absence of official protocols in handling digital privacy data, the proposed scheme was used to protect the unrelated private data stored on the email server. Encryption was firstly adopted at the data acquisition stage to protect the data content. At the same time, the keywords that were related to the case investigation were indexed. After applying the encryption, the encryption key was kept by the server owner so as to prevent any unauthorized access of data by the digital investigator.

The encryption process on the email server took approximately two hours. Thereafter, the examination was conducted on this protected image using specific keywords inputs. If there was a keyword hit on the data, the respective data area, i.e. email, would be decrypted. That allowed the investigator to conduct further computer forensic examination.

In contrast with the traditional approach which solely relied on the integrity of the investigator in protecting the private data, this methodology had practically enhanced the confidence of the data owner regarding the investigation and at the same time maintained the overall effectiveness of the investigation as a whole. Though the whole process was slower than the traditional approach, the ultimate goal of protecting digital data privacy during digital investigation process was achieved. Subject to further review on other practical case scenarios, this approach is expected to be appropriate for dealing with cases which involve sensitive private data.

VIII. Conclusions

In this paper, we address an important, but not adequately addressed in the community, issue for protecting digital privacy data information during forensic investigation. We highlighted the differences of digital privacy information and physical privacy information protection and discussed the difficulties of handling digital privacy data. We reviewed the current practices for computer forensic examination and proposed the way in using encryption in protecting privacy data in the context of forensic examination.

Based on a simulated scenario, a study on the proposed scheme was carried out to verify the feasibility of the approach as well as to understand the effect of encryption to the data content. The proposed approach, of course, is not the only solution to the problem. Finding a better scheme and procedure to solve this emerging problem is always desirable. We hope that this paper can catch the attention of the community to help developing a forensically sound procedure to solve this problem.

References

[1] B.Carrier, E.H.Spafford. Getting Physical with the Digital Investigation Process, International Journal on Digital Evidence, Fall 2003, Volume 2, Issue 2.

[2] B.Markham. Data Classification and Privacy: A foundation for compliance, http://www.oit.umd.edu/Publications/Data_Classification_Presentation_022908.pdf

[3] ACPO Guideline on Good Practice Guide for Computer based Electronic Evidence, http://cryptome.org/acpo-guide.htm

[4] DFRWS, Report from the First Digital Forensic Research Workshop. DTR-T001-01 FINAL A Road Map for Digital Forensic Research. Final version, November 6, 2001.

[5] M.Geiger and L.Cranor. Counter-Forensic Privacy Tools – A Forensic evaluation. CMU-ISRI-05-119, Institute of Software Research, International, Carnegie Mellon University, June 2005.

[6] Data Protection Principles of the United Kingdom, http://www.dfpni.gov.uk/section_10_data_protection_principles.pdf

[7] P.Stahlberg, G.Miklau and B.N.Levine. Threats to Privacy in the Forensic Analysis of Database Systems, SIGMOD 07, Beijing, China.

[8] Orin S. Kerr, Searches and Seizures in a Digital World. Harvard Law Review, Vol. 119, 2006; GWU Law School Public Law Research Paper No. 135.
[9] B.Carrier. File System Forensic Analysis. Addison-Wesley Professional, 2005.
[10] E.Casey. Digital Evidence and Computer Crime. Elsevier, 2nd edition, 2004.
[11] S.L.Garfinkel and A.Shelat. Remembrance of data passed: A study of disk sanitization practices. IEEE Security and Privacy, Jan/Feb 2003.
[12] Automated Data Processing Security Manual, Department of Defense Manual, DoD 5200.28-M, June 1979.
[13] Care and Handling of Computer Magnetic Storage Media, Department of Commerce, National Bureau of Standards Special Publication 500-101, June 1983.
[14] Department of the Navy Automated Data Processing Security Program, Chief of Naval Operations Instruction, OPNAVINST 5239.1A, 1982.
[15] Remanence Security, Air Force Systems Security Instruction, AFSSI 5020, April 1991.
[16] National Industrial Security Manual for Safeguarding Classified Information, Department of Defense Manual, DoD 5220.22-M, June 1987.
[17] KATZ v. UNITED STATES, 389 U.S. 347 (1967)
[18] Privacy Amendment (Private Sector) Act 2000, Australia
[19] Personal Data (Privacy) Ordinance, Chapter 486, Laws of Hong Kong
[20] R.Ieong, “How to Balance Privilege and Digital Forensics Investigation”, The Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Kaohsiung, Taiwan, November 2007.
[21] R.Ieong, and H. C. Leung, “Deriving case-specific Live Forensics Investigation procedures from FORZA”, Proceedings of the 2007 ACM symposium on Applied computing, Seoul, Korea, March 2007.
[22] M.Geiger and L. Cranor, “Scrubbing stubborn data: An evaluation of computer forensic privacy tools”, IEEE Security and Privacy Magazine, 4(5):16-25, 2006
[23] Tax website shut down as memory stick with secret personal data of 12million is found in a pub car park, Mail Online, http://www.dailymail.co.uk/news/article-1082402/Tax-website-shut-memory-stick-secret-personal-data-12million-pub-car-park.html
[24] D. Boneh and R. Lipton, “A revocable backup system”, USENIX Security Symposium, pp 91-96, 1996
[25] D. Boneh, G. Di Crescenzo, R. Ostrovsky and G. Persiano. Searchable public key encryption. In Advances in Cryptology – Eurocrypt ‘04.
[26] P. Golle, J. Staddon and B. Waters, "Secure Conjunctive keyword search over encrypted data," ACNS 2004, LNCS 3089, pp. 31-45, Springer-Verlag, 2004.
[27] Dawn Xiaodong Song, David Wagner, Adrian Perrig. Practical Techniques for Searches on Encrypted Data. In Proceedings of IEEE Symposium on Security and Privacy, 2000.
[28] Eu-Jin Goh, “Secure Indexes”, Cryptology ePrint Archive, Report 2003/216, 2003 (http://eprint.iacr.org/2003/216/).
[29] D.J. Park, K. Kim and P.J. Lee. Public Key Encryption with Conjunctive Field Keyword Search. In Information Security Applications. 2004.
[30] Oracle Corporation. Database Encryption in Oracle9i, 2001.
[31] IBM Data Encryption for IMS and DB2 Databases, Version 1.1, 2003.G