# An Investigation into Reducing Third Party Privacy Breaches During the Investigation of Cybercrime

Wynand JC van Staden
School of Computing
University of South Africa, Science Campus
Johannesburg
South Africa
wvs@wvs.za.net

***Abstract*—In this article we continue previous work in which a framework for preventing or limiting a privacy breach of a third party during the investigation of cybercrime. The investigations may be conducted internally (by the enterprise), or externally (by a third party, or a law enforcement agency) depending on the jurisdiction and context of the case. In many cases, an enterprise will conduct an internal investigation against some allegation of wrongdoing by an employee, or a client. In these cases maintaining the privacy promise made to other clients or customers is an ideal that the enterprise may wish to honour, especially if the image or brand of the enterprise may be impacted when the details of the process followed during the investigation becomes clear. The article reports on the results of the implementation of the privacy breach detection – it also includes lessons learned, and proposes further steps for refining the breach detection techniques and methods for future digital forensic investigation.**

***Index Terms*—Privacy, Digital Forensics, Privacy Breach, Third Party Privacy, Cybercrime**

## I. Introduction

A Digital Forensic (DF) investigation relies on the ability of the investigator to build a coherent and consistent description of events of the incident being investigated. In order to accomplish this, the investigator should have access to all the relevant information that may provide context in the case. However, the investigator may be granted access to more information than is, strictly speaking, needed. For example, an investigator searching for proof of pornographic images containing children may conduct a legitimate search for all images on a storage medium that was shared by more than one person. A subsequent search may return all images on the disk – thus the potential for a privacy breach exists – the event that a third party, not culpable in the actions leading to the investigation, may be 'investigated' (we define this as a Third Party Privacy Breach (TPPB)). Such privacy breaches may not be considered harmful, however, it could reveal a person's political affiliation, detail on one's social circles and so on – the detail of which could place the third party in a compromised position.

During post-mortem analysis of data, the investigator uses a plethora of tools in order to sift through large volumes of data – the tools also provide clues to the documents and artefacts that may be of interest. A common tool used is a regular expression parser and searcher, which will scan a collection of documents for keywords. The investigator may then (painstakingly) examine the documents in order to determine their relevance and to build a case. This requires that all results returned will be considered, and data relevant to the investigation will be retained, and non-relevant data will be discarded.

In previous work [1], we proposed a method for limiting the possibility of a privacy breach through the use of information retrieval techniques such as clustering, and diversity index creation. The principle behind the proposal is to release the results of a search only if it can be determined that the query being posed to the system is not considered 'too wide' or 'unspecific'. A query is deemed too wide if it may reveal information on third parties. A focussed query on the other hand returns results that fit close together, and thus has a smaller chance of revealing information on the third party.

In this paper we continue that work by presenting the results of implementing the Information Retrieval (IR) system on the well known (and researched) ENRON email corpus. The corpus contains real world data which consists of communication relating to the core business of the ENRON corporation, as well as numerous emails of a personal nature. Our intent was to determine how well the proposed system would aid in filtering out searches that would be considered possible privacy breaches.

### A. Contribution

This paper contributes to the general area of digital forensic knowledge in the following way: the application of well known information retrieval techniques to DF investigations has been proposed elsewhere [2], [3]. By extending search techniques to include privacy considerations will aid in the prevention of TPPBs, thereby ensuring that the right to privacy of persons are protected even when an investigation takes place. We have implemented a previously proposed framework using the techniques and a generally available corpus of real world data to examine the potential for privacy breaches and the mitigation of this risk.

Our results provide insight into the complexity of the analysis of searches using IR techniques for privacy protection and provides clues and avenues for future research in this area. The questions and areas of difficulty identified in this paper

978-1-4799-3383-9/14/$31.00 ©2014 IEEE

will be used to guide future research with the specific intent of finding better ways of limiting privacy breach exposure during a digital forensic investigation. Limiting privacy risk in such a way is becoming more and more relevant when one takes the recent reports of government-sanctioned information gathering into account: what once was an implicit trust of those in power to behave responsibly with Personal Identifiable Information (PII) has now become a matter of public interest and action.

### B. *Structure of the paper*

The rest of the paper is structured as follows: section II provides relevant background and related material, section III provides a description of the implementation that was done (the artefact). Section IV provides results that were forthcoming from the implemented search prototype. Section V provides a summary of the insights gathered after analysing the results, and finally section VI provides concluding remarks, and proposes avenues for future research.

## II. Background and related work

The work presented in this paper covers two areas of research: digital forensic investigation of cybercrime, and privacy. This section provides a brief background on each, as well as a short description of our previous proposal on TPPBs in order to provide context.

### A. *Digital Forensics*

Digital forensic science is "the use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose facilitating or furthering the reconstruction of events found to be criminal or helping to anticipate unauthorised actions shown to be disruptive to planned operations." [4]. The preservation and collection of of digital evidence results in the seizure of any computer equipment that has been identified as sources of digital evidence [5], [6]. This paper focuses on the instances in which storage media (hard disks, DVDs, flash-drives, and so on) is seized for analysis.

The process followed (as described above) is not limited to governmental institutions and it is common for large enterprises to house internal investigative units which will conduct investigations and that may hand the investigation to a governmental law enforcement agency. On the other hand, the enterprise may not have the expertise or experience needed to conduct the investigation, and the incident may be outrightly handed over.

Analysis of the storage media seized is referred to as a post-mortem analysis, and requires the scrutinizing of both used and non-used areas of non-volatile storage. Used areas contain data that is managed by the file-system (including areas that are reserved for usage but that is not associated with any container managed by the operating system – referred to as slack-space), and non-used areas contain residual data (data that was used at some point, but of which the container was removed). Used storage typically consist of directories and files that contain data. Files are classified according to the type of data they contain and may be image data (binary), plain text data, Hyper-text Markup Language (HTML), compressed data, or encrypted data. Encrypted data is meaningless unless the key used to decrypt the data is known to the investigator, and as such may hinder an investigation. Getting access to the relevant keys is normally done by petitioning the government to exert force on the key-holder to relinquish the key. The rest of the paper therefore assumes that all data under consideration has been decrypted.

In the case of multi-user systems, the operating system will store the data for each user separately, and access to the data is enforced through the classic information security mechanisms such as Access Control Lists (ACLs), an Access Control Matrix (ACM), or tokens [7]. However, most current operating systems provide unrestricted access to data on the system through a super-user or administrator account (even if this were not the case, it is a fairly trivial task to construct a program that will be able to bypass the operating system and provide unrestricted access to the data). In the event of an investigation the investigator needs unrestricted access to the data being analysed. This means that the barriers that existed through the access control mechanisms listed above no longer applies, and that the investigator has the same unrestricted access to the data as the super-user.

Searches for data that may be of interest to the investigator requires high recall [3] and is commonly done by running pattern matching searches on the data on the storage media. These searches reveal (in the case of used space) all the files that contain data that match the pattern, and in the case of non-used space the blocks on disk that contain matching patterns. The analyst then constructs evidence by carefully examining the results and including files and blocks that may support the case and discarding the files and blocks that match the pattern but that are unrelated to the case.

It is this unfettered access to information that raises the question of privacy. An investigator on a multi-user system may (as a result of their search query) be presented with files that are unrelated to the investigation, but are returned because they contain data that match the pattern specified by the investigator. We consider the ability of the investigator to view these files as a potential TPPB.

### B. *Privacy*

Privacy is defined as the right to information self determination [8], [9], or the ability to state who has information about you, what they can do with it, and with whom they can share it. The foundations of privacy are well established [10], [11], and the ability of a system to offer privacy protection has been well researched, and an architecture for such systems has been proposed [12]. However, despite this fact, one can easily find newspaper articles on the relative ease with which privacy breaches take place. The PRIME [13] project (and PRIMElife) attempts to put a framework in place in which systems adhere to the promises they make regarding privacy.

Privacy is a big concern for a society in which data gathering and analysis is becoming easier as technology improves. The manner in which modern society interacts through social media also provides a means for data gathering and analysis which poses serious threat to privacy. And this expectation created by information self determination may have a serious impact on how investigations into computer crime is conducted – as a user of a multi-user computer system one does expect that PII will be kept private. However, as indicated above, the investigator may be granted unlimited access to all the data on the system. Systems that state that they take privacy of their users into account creates an expectation of privacy, which may cause problems for the data controller [1] if a breach occurs.

In our previous work [1] we proposed a method for limiting the risk of a privacy breach by avoiding searches for patterns which yields a search result that may have a privacy breach. The proposal used a search filter which classifies results based on an automated topic clustering algorithm [14] and a diversity rating for the results based on Shannon's diversity index [15].

Beebe and Clark [3] proposed a thematic clustering technique for search results on order to cluster similar results based on Kohonen's Self-Organising Maps (SOMs) [16], Fei and Olivier [17] used SOMs to identify anomalies in search results. Both of these techniques attempt to cluster data in a way that eases the burden on the investigator. The primary difference between the approach used in this paper and the mentioned work, is that we require the search results to cluster documents that are similar purely to determine if they cover different topics.

A too diverse set of topics would indicate a potential for a privacy breach – the reasoning is that a set of documents that are closely related better reflects a focussed query, and the results that are returned thus poses less of a privacy breach risk. In essence the system rewards patterns or queries that yield results that are closely related. We proposed a framework in which the risk of a TPPB would be reduced. This framework consisted of an indexing system, which is used during the search, and a filtering system which would calculate a diversity index on the resulting corpus. An ideal index for a search query is calculated and if the resulting corpus deviated too much from the ideal index, the query is denied (no results are returned). This forces the investigator to narrow the query using more specific search terms. To avoid situations where the investigator is stymied by the filtering system, the notion of a threshold was introduced which could be shifted by the investigator to allow diverse queries to yield results. The shifting of the threshold is recorded in an audit log for later examination if it should happen that a breach of privacy is reported.

In the following section we provide detail on the implementation of our experiment.

## III. Implementation Detail

In this section we provide detail on the construction of the proof of concept implementation of search and filter portion proposed framework. The proof of concept makes use of standard information retrieval techniques [18], run on the Enron email corpus[2] . The corpus itself consists of roughly 517000 email messages from 151 employees. All attachments were removed from the emails and the messages themselves are listed in files in RFC4155 (mbox) format [19].

We created an inverted term index on the entire Enron corpus using the Single-pass In Memory Indexing (SPIMI) method in Python. Extracted terms were stemmed using the Porter [20] stemmer. Our tokenizer for the email content delivered just over 780 000 unique terms which includes phone numbers, 16 to 19 digit numbers resembling credit card numbers, email addresses, web addresses, and file-names. The resulting index was indexed using a simple BTree to speed up search results. The entire scanning and indexing process of the 3.6Gb corpus took a little over 30 minutes on our modest hardware[3] (given the simplicity of our implementation a more robust implementation should be significantly faster).

Queries were stemmed, and the indexes used to find relevant mails (mbox files). Using the N-Gram based technique as proposed by Cavnar et al [14] the *distance* between the mails in the results was calculated. These distances was stored for later use (in order to speed-up subsequent queries). Once distances were calculated we employed a single linkage clustering method for clustering emails based on topic. Once clustering was complete, we used the Shannon entropy index to calculate the diversity index on the corpus, and we calculated an ideal diversity index for the corpus. The ideal index or *diversity norm* is an diversity indicator for a result corpus containing a single cluster with all the mails contained in it (or equivalently a small number of clusters with mails spread evenly between them). The diversity norm is thus the ideal, and the actual diversity index is compared to it. To group related emails, we employed a single-linkage clustering scheme as proposed in our original paper. The resulting clustered corpus was then given a diversity norm index as well as a diversity index. The diversity norm being the ideal situation in which a query is focussed enough to return only documents that are related in topic, and the diversity index is an indicator of the diversity of topics being covered in the result corpus. If the norm and the diversity index differed too much (past a threshold) no results were returned.

We made no attempt to try and find 'culprits' to the original Enron saga in our investigation: our primary aim was to find indications of privacy leaks using regular techniques that a forensic investigator would employ, and to see what types of results were returned. Several search terms related to an Energy supplier's business were used, and we generated reports indicating the number of clusters, the diversity norm,

[1] The enterprise that stores the data

[2] Available for download from http://www-2.cs.cmu.edu/ enron/

[3] Intel I5 architecture, with 16Gb of Random Access Memory (RAM), and a 7200rpm SATA disk drive.

and the diversity index. To further aid in our investigation we used the Natural Language Tool kit (NLTK)[4] to generate collocation and concordance reports for the documents. Broad searches (few search terms) in many cases resulted in a large number of mails in the search results (see table I).

| Search phrase used | Number of emails returned |
|---|---|
| connection | 13569 |
| hook up | 1709 |
| pipeline | 15940 |
| chief executive officer | 5330 |
| federal energy regulatory commission | 4230 |

TABLE I
BROAD SEARCH RESULTS

We extended our search to include words that may form part of business vernacular, or social context. These included words like ‘graduate’, ‘student’, ‘loan’, ‘weekend’, and so on. A standard expression search applied to the Enron mail corpus produces some interesting results. For the person searching for emails relating to specific topics (using keywords) the slightly larger than 517000 email corpus provides some interesting results. For example the phrase ’hook up’, a term used in the Enron corpus to refer to a new gas or power connection (or an ADSL connection – Enron branched into internet connectivity as well) returns several emails referring to new business proposals and requests for meetings to discuss the potential sale of new connections to energy providers, or clients.

However, the term ’hook up’ may also refer informally to a social get-together. The implication here is that an ’innocent’ search for emails relating to new business, when investigating persons involved in fraudulent sales activity (for example) will also return detail on other member’s of staff’s personal life – something they might not want to be exposed during the investigation, but will be because the investigation relies on the availability of data in order to search.

A keyword search for ’graduate’ yielded emails on an intern programme that was run at Enron, as well as discussions about repayment of student loans, as well as some employees’ plans to enrol for graduate school. In many of these cases, the emails that turn up are in the mail-collection of employees that are in different departments to the emails of those who are of interest (those who are involved in the sales of Enron products). They appear as a person of interest purely because they used a word that has a bearing to the one the investigator is using, but the context differs – the topic of the email is different from the ones that are being targeted.

We could consider the emails harmless in the face of a legitimate criminal or civil investigation – if trust were absolute we would be assured that this personal information would never leak, however, past experience should convince us otherwise. And in the event the information is leaked, we could argue that in spite of this a greater good has been served, however, we would be ignoring the principle of privacy, as well as an ethical problem: in that the liberties of a few cannot be ignored in favour of an aggregate good being served.

In the event of an internal investigation, an enterprise may uncover information about an employee that would constitute information asymmetry to the disadvantage of the employee – a situation which is similar to constant email monitoring in most respects.

## IV. RESULTS

In this section we present some results and draw some conclusions on the value of the results.

Our searches indicated that ’innocent’ keyword searches reveals privacy information. This is definitely not always the case, especially if terminology that is extremely narrow in scope was used in the search. However, when searching for keywords in a broader context,we found that search results contained email detailing personal encounters, email with humorous content (but which contained explicit depictions). Email with explicit content in and of itself may not be considered private, but it is exactly these emails that are could be used by media to label persons during a trail. These labels may change the perception others have of the persons whose information is leaked and we may consider that a privacy leak.

Furthermore, searches which included the keywords ’graduate’ or ’loan’ returned references to social security numbers, and personal identification numbers. It is thus clear that a search with high precision may result in a privacy breach.

Clustering the search results using the proposed method produced the following results: firstly, for broad searches such as ’hook up’, ’meeting’, or ’request’, the resulting clustered corpus returned had an extremely high diversity index, as we expected. Narrowing the searches resulted in a sharp drop in the diversity count. Tables II, and III summarise our findings.

| Search phrase used | Number of emails returned |
|---|---|
| hook up | 1709 |
| hook up request | 434 |
| hook up new request gas pipeline | 219 |

TABLE II
BROAD KEYWORD SEARCH WITH SPECIFIC INTENT

| Search phrase used | Diversity norm | Diversity Index | Nr of clusters |
|---|---|---|---|
| hook up | 10.7 | 53.9 | 37 |
| hook up request | 8.7 | 48.6 | 32 |
| hook up new request gas pipeline | 7.7 | 48.4 | 30 |

TABLE III
REDUCTION IN DIVERSITY FOR NARROWING SEARCH

What becomes apparent is: firstly, in many cases there is a drop in the number of files returned as more keywords are added (thus the query starts focussing on a particular context). Secondly, depending on the query term, the diversity index and clusters starts stabilising. However, the stabilisation is far from ideal: in the example presented in table III the query would

[4] www.nltk.org

still be deemed too wide and filtered out. However, a manual inspection of the resulting corpus revealed that the emails were either news-feed summaries, correspondence regarding new business, or new request for gas pipeline connections ("hook-ups"). Thus, we have succeeded in filtering out emails that may be included incidentally based on a raw keyword search, however, the more focussed query was still deemed too wide because of the blind clustering technique.

To enhance the clustering we implemented the BM25 [21] ranking algorithm on the search results, and clustered the top 100 documents. The diversity index was reduced, but did not perform significantly better than the blind clustering.

The following section discusses some of the lessons learned during our investigation into the techniques used in this paper.

## V. Lessons Learned

In this section we share some insights and lessons learned in the approach taken as applied to real world data which most reflects the intent of the original proposal: a multi-user system in which certain users were to be investigated for a certain crime and in which a particular approach to finding relevant emails would reveal information of a personal nature about third parties.

Firstly, even our rudimentary implementation performed well and indexed the 517 000 emails in a negligible amount of time, thus not significantly interfering with an investigation. Searches using the inverted term index and BTree index on the inverted index completed in sub-second times.

Secondly, and arguably the most important lesson learned from the experiment is that the critical factor in reducing privacy breach exposure using a blind topic categoriser is at best difficult – we found in too many cases that the number of clusters returned by the clustering component indicated a too diverse set of topics. Manual inspection of the returned corpus revealed that many of the clusters were related, however, the single-linkage clustering technique clustered them correctly based on the initial protocol devised. Based on this it is apparent that an agglomerative clustering algorithm may perform better. However, determining the level at which the re-clustering should take place will need further investigation.

Thirdly, the BM25 ranking algorithm was also a fairly good indicator of the number of clusters in the corpus. We found invariably that a wide range in BM25 scores indicated a large number of clusters. Ranks with little difference in them would typically correspond to a cluster as returned by the clustering algorithm. This indicates a strong correlation between automated blind topic categorisation and ranking based on BM25.

Finally, the reports generated on the search queries provided strong evidence that collocation analysis as reported by Wanner and Ramos [22] would provide a better document clustering technique during the investigation.

The following section provides concluding remarks.

## VI. Conclusion

This article presented an investigation into the application of a framework for preventing a TPPB during a digital forensic investigation.

As details of the Enron saga was released, it was shown that many people not involved in the insider trading and fraudulent activities had some of their personal information released. Our work illustrates that privacy breaches for these third parties may have unintentionally happened.

We presented a proof of concept of the search and filter portion of a proposed framework for preventing TPPB done previously, and showed how this application to the Enron mail corpus could have restricted the amount of embarrassment felt by third parties. In situations were internal investigations are normally conducted in civil liability cases, the application of the tool could be used to avoid leakage of private information, thereby protecting the good standing of the enterprise in the community.

We also showed that the clustering technique, although useful requires refinement – our initial study shows that a blind clustering does not provide a suitable diversity index – the nature of the emails and terminology places even a few emails in diverse clusters. Initial good indicators are that collocation phrases of the texts may be a good indicator of the topic of the email. We are in essence doing a post-application of mail-folder management, however, not from the user's point of view.

Another possible approach is classification using machine learning techniques such as naive Bayesian classification. However, we purposefully avoided using guided training for two reasons. Firstly, since the enterprise's email will be specific to their industry, a set of training data may not be applicable to the email corpus, resulting in a low revel of recall, and the results can therefore not be trusted. Secondly, guided training in such a specific setting then means that someone will have to manually examine existing emails, and classify them so that they can be used as training data. This again poses a risk to privacy. The use of Natural Language Processing (NLP) techniques to detect context may be of value and we have started investigating this possibility, and will report on it elsewhere.

## References

[1] W. J. van Staden, "Third Party Privacy and the Investigation Of Cyber-crime," in *Advances in Digital Forensics IX*, G. Peterson and S. Shenoi, Eds. Orlando, Florida, USA: Springer, 2013.

[2] N. Beebe and J. Clark, "Dealing with Terabyte Data Sets in Digital Investigations," in *Advances in Digital Forensics*. Springer US, 2005, vol. 194, ch. IFIP — The International Federation for Information Processing, pp. 3–16.

[3] N. L. Beebe and J. G. Clark, "Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results," *Digital investigation*, vol. 4, pp. 49–54, 2007.

[4] G. Palmer, "A Road Map for Digital Forensic Research," DFRWS, Utica, NY, Tech. Rep., 2001.

[5] Technical Working Group for the Examination of Digital Evidence, *Forensic Examination of Digital Evidence: A guide for law enforcement*, 2002.

[6] R. Nolan, C. O'Sulivan, J. Branson, and C. Waits, "First Responders Guide to Computer Forensics," Carnegie Mellon Software Engineering Institute, Pittsburgh, Pennsylvania, Tech. Rep., March 2005.

[7] C. P. Pfleeger and S. L. Pfleeger, *Security in Computing*, Fourth ed. Prentice Hall, 2012.

[8] S. Fischer-Hübner, *IT security and privacy: Design and use of privacy enhancing security mechanisms*. Springer-Verlag, 2001.

[9] W. v. Staden and M. S. Olivier, "On Compound Purposes and Compound Reasons for Enabling Privacy," *J. UCS*, vol. 17, no. 3, pp. 426–450, 2011.

[10] S. D. Warren and L. D. Brandeis, "The right to privacy," *Harvard Law Review*, pp. 193–220, 1889.

[11] "OECD Guidelines on the Protection of Privacy and transborder Flows of Personal Data," Organisation for Economic Cooperation and Development, Tech. Rep., 1980.

[12] R. Hes and J. Borking, Eds., *Background Studies and Invesitigations 11: The road to anonimity*, Revised ed. Registratiekamer, Den Haag: Dutch DPA, August 2000.

[13] J. Camenisch, A. Shelat, D. Sommer, S. Fischer-Hübner, M. Jansen, H. Kraseman, R. Leenes, and J. Tseng, "Privacy and Identity Management for Everyone," in *DIM'05*. Fairfax, Virginia, USA: ACM, November 2005.

[14] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorisation," in *SDAIR-94, Third Annual Symposium on Document Analysis and Information Retrieval*, 1994, pp. 161–175.

[15] C. E. Shannon, "A Mathematical Theory of Communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[16] T. Kohonen, "The Self Organising Map," in *IEEE*. IEEE, 1990, pp. 1464–1480.

[17] B. K. L. Fei, J. H. P. Eloff, M. S. Olivier, and H. S. Venter, "The use of self-organising maps for anomalous behaviour detection in a digital investigation." *Forensic Sci. Int.*, vol. 162, no. 1-3, pp. 33–7, 2006.

[18] C. D. Manning, P. Raghavan, and H. Scütze, *An Introduction to Information Retrieval*. England: Cambridge University Press, 2009.

[19] E. A. Hall, "The application/mbox Media Type," Electronically, September 2005, http://datatracker.ietf.org/doc/rfc4155/.

[20] M. F. Porter, "An Algorithm for Suffix Stripping," K. Sparck Jones and P. Willett, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, ch. Readings in Information Retrieval, pp. 313–316. [Online]. Available: http://dl.acm.org/citation.cfm?id=275537.275705

[21] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," 1996, pp. 109–126.

[22] L. Wanner and M. A. Ramos, "Local Document Relevance Clustering in IR Using Collocation Information," in *The fifth international conference on Language Resources and Evaluation*, 2006.