A Suspect-Oriented Intelligent and Automated Computer Forensic Analysis

M. Al Fahdi[1], N.L. Clarke[1,2], F. Li[1], S.M. Furnell[1,2,3]

[1]School of Computing, Electronics and Mathematics, Plymouth University, UK
[2]Security Research Institute, Edith Cowan University, Western Australia
[3]Centre for Research in Information and Cyber Security, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

info@cscan.org

Abstract.

Computer forensics faces a range of challenges due to the widespread use of computing technologies. Examples include the increasing volume of data and devices that need to be analysed in any single case, differing platforms, use of encryption and new technology paradigms (such as cloud computing and the Internet of Things). Automation within forensic tools exists, but only to perform very simple tasks, such as data carving and file signature analysis. Investigators are responsible for undertaking the cognitively challenging and time-consuming process of identifying relevant artefacts. Due to the volume of cyber-dependent (e.g., malware and hacking) and cyber-enabled (e.g., fraud and online harassment) crimes, this results in a large backlog of cases. With the aim of speeding up the analysis process, this paper investigates the role that unsupervised pattern recognition can have in identifying notable artefacts. A study utilising the Self-Organising Map (SOM) to automatically cluster notable artefacts was devised using a series of four cases. Several SOMs were created – a file list SOM containing the metadata of files based upon the file system, and a series of application level SOMs based upon metadata extracted from files themselves (e.g., EXIF data extracted from JPEGs and email metadata extracted from email files). A total of 275 sets of experiments were conducted to determine the viability of clustering across a range of network configurations. The results reveal that more than 93.5% of notable artefacts were grouped within the rank-five clusters in all four cases. The best performance was achieved by using a 10x10 SOM where all notables were clustered in a single cell with only 1.6% of the non-notable artefacts (noise) being present, highlighting that SOM-based analysis does have the potential to cluster notable versus noise files to a degree that would significantly reduce the investigation time. Whilst clustering has proven to be successful, operationalizing it is still a challenge (for example, how to identify the cluster containing the largest proportion of notables within the case). The paper continues to propose a process that capitalises upon SOM and other parameters such as the timeline to identify notable artefacts whilst minimising noise files. Overall, based solely upon unsupervised learning, the approach is able to achieve a recall rate of up to 93%.

Keywords: SOM, digital forensics, automation, clustering, self-organising map, cybercrime

1 Introduction

Over the last 15 years, computing technologies have experienced significant change in terms of variety of devices (e.g., computers, smartphones and tablets), data capacity (e.g., storing up to and beyond 2 Terabytes (TB) of data), functionality (e.g., office, web browsing and mobile apps), and the number of users. Indeed, the use of computing devices has integrated into every aspect of daily life such as email, banking, entertainment, shopping, and micro-payments. Unfortunately, in parallel with this, the types and sophistication of computer assisted cybercrimes have also grown significantly, from the traditional child pornography, fraud, and money laundering to carefully planned cyberattacks (e.g., government espionage, cyber warfare, and identity theft). Inevitably, the consequence of these cybercrimes can be severe. For UK businesses alone, cyberattacks are claimed to have cost £34 billion in lost revenue in 2014 (Veracode, 2015).

Digital Forensics has become an invaluable tool in the identification of cybercriminal activities due to its ability to extract valuable information and evidence from computing devices in a legally acceptable manner (Casey, 2010). As a result, it has been widely used by law enforcement agencies and organisations to track and investigate computer-assisted and cybercriminal activities (Inforsecusa, 2011; Brainz, 2014; RCFL, 2014).

However, digital forensics experiences growing challenges from several aspects, including the growing size of data storage, the prevalence of embedded flash storage, the need to analyse multiple devices, the use of encryption, and the popular usage of cloud computing (Casey and Stellatos, 2008; Garfinkel, 2010). Statistics from the FBI's Regional Computer Forensics Laboratory (RCFL) show that they had processed 5,973 TBs of data from 7,273 examinations in 2013 – a 40% increase in comparison with 2011 (FBI, 2013). Despite their effort, an audit report of the Office of the Inspector General U.S. Department of Justice highlights that a backlog of 1,566 outstanding cases existed, 57% of which had waited between 91 days to over 2 years (Office of the Inspector General, 2015). Unfortunately, the consequence of such backlogs could cause a number of implications, both legal and personal.

In order to reduce the overall examination time, many forensic tools have been developed both commercially or under open source licence agreements, such as EnCase (Guidance Software, 2015), Forensics Toolkit (FTK) (AccessData, 2015), P2 Commander (Paraben Corporation, 2015), Autopsy (Carrier, 2015), HELIX3 (e-fense, 2014), and Free Hex Editor Neo (HHD Software, 2015). The majority provide the "Push-Button Forensics" facility to automate several key procedures of the forensic process, including preservation, collection, and presentation. Despite the assistance of these tools, digital evidence examiners still have to manually analyse the data (e.g., documents, emails, and internet history) contained on the image to find potential evidence; however, this process is time consuming and prone to human-error. Also, it is the responsibility of the investigator to cognitively analyse the data and understand

the inter-relationships that exist between artefacts. On cases with a growing volume of data, this places an ever-increasing burden upon the investigator. Indeed, this has led many law enforcement agencies to strategically change their approach away from the '*gold standard*' (analysing all files to ensure nothing is overlooked) to '*intelligence-based*', where a subset of files are analysed dependent upon the intelligence provided to the investigator (Lawton et al, 2014). It is no longer about finding every piece of evidence but rather sufficient evidence to determine innocence or guilt. To this end, this paper describes a novel analysis approach that utilises the Self-Organising Map (SOM) technique to automatically group artefacts of interest together, enabling investigators to focus specifically on notable files (i.e., those that are relevant to the case) and hence reduce the time spent on analysing irrelevant files. The approach is based upon utilising the metadata from a variety of sources, such as the file system (e.g. pathname, file type, and Modification, Access and Creation (MAC) timestamps) and email (e.g. to, from, and attachment present) as an input into the SOM clustering. An experiment is presented to illustrate whether clustering is a viable approach to identifying notable artefacts.

The remainder of the paper is structured as follows: Section 2 presents the existing work surrounding the use of SOM clustering with respect to digital forensics. Section 3 describes the datasets that were utilised in the experiment, with Section 4 presenting the experimental results of the SOM study. Section 5 presents a novel process that applies SOM in practice and presents an evaluation of the approach using the aforementioned datasets. A comprehensive discussion on the impact of the results in practice is presented in Section 6, prior to the conclusion and future work.

2 Related Work

A SOM is a neural network that produces a mapping from the high dimensional input data into a regular two dimensional array of nodes based upon their similarity (Kohonen, 1998). Due to its competitive learning nature, SOM can automatically classify the input data without any supervision. Since its invention, SOM has been extensively used in many computer security related fields, including intrusion detection, biometrics, and wireless security (Feyereisl and Aickelin, 2009). The use of SOM within the digital forensic domain can be traced back in the early 2000s, where police were able to link records of serious sexual attacks together (Adderley and Musgrove, 2001). Since then, a number of studies were devised to investigate the ability of SOM for digital forensic investigations.

Fei et al (2005) and Fei et al (2006) explored the use of SOM as a supporting technique to interpret and analyse data generated by computer forensic tools in a visualised manner. In their studies, a public dataset containing 2,640 graphical images was utilised; each image contained four features: the file name, extension, creation time, and creation date. SOM clustered the data after being manually enumerated, producing various two dimensional maps. These visualisations enabled digital evidence examiners to locate interesting information in a more efficient and accurate manner.

However, experimental results were not presented in detail to highlight the efficiency and accuracy of their proposed approach.

With the purpose of improving the result of text-based searches, Beebe and Clark (2007) proposed a novel method that utilised SOM to post-retrieval cluster text string search results within a forensic image. In order to test their hypothesis, a software tool (named "Grouper") was developed. Grouper was able to perform a number of activities, including data preparation and SOM clustering. Experimentally, two datasets were utilised: one was a real-world divorce case and the other was an artificially created murder case; their image sizes were 40 and 10 Gigabytes (GB) respectively. Results demonstrated that the approach can be used to reduce the human analytical time by around 80% despite additional computer processing time being required (Beebe et al, 2011).

Kayacik and Zincir-Heywood (2006) created a topological model of known attacks for forensic analysis of anomalous network traffic by employing the SOM algorithm. Their model was tested by using the KDD 99 intrusion detection dataset. The results of their empirical study show that attacks can be successfully grouped by SOM with an overall high accuracy (i.e., 89.8%). Also, they suggested that the model can be utilised for analysing new attacks or suspicious network behaviour.

Similarly, Palomo et al (2011) focussed upon the analysis and visualisation of network traffic data via the use of SOM to identify abnormal behaviour or intrusions. For their experiment, a dataset with 150,871 packet samples was created by monitoring a university network via WireShark during a four-day period; each sample contained nine features, including the IP addresses of source and destination, port numbers, protocol type, date and time stamps, and packet length. The data was clustered using SOM with various network configurations (e.g., 3x3 and 5x5 network sizes). Their experimental results demonstrate that suspicious network traffic was identified by SOM providing vital information for network forensic examiners.

Wang et al (2015) proposed a graphical model to analyse the relationship between criminals through SOM visual analytics. Their model was evaluated using a dataset with 16,383 features of 16 suspects. Within the model, SOM was used to reduce features and provide a visual aid to investigators for better understanding of suspect's activities. According to their experimental results, the proposed model can offer assistance for a more efficient forensic analysis. Nonetheless, the degree of assistance to which the model was able to offer was not clearly provided.

As illustrated above, SOM has been used in several digital forensic domains, including image analysis, network forensics, and text-based searching. The results suggest that SOM can be used successfully to assist the forensic examiners, such as in the visualisation of artefacts and the reduction of human analytical time. Nevertheless, the ability to use SOM to analyse a forensic image specifically tasked with identifying notable files using metadata extracted at the file system and application levels has never been undertaken.

3 Datasets

In order to investigate the ability of using SOM to cluster notable artefacts, four forensic cases were utilised: two public and two private. Whilst it would have been useful to utilise a larger number of cases, the availability of these (for obvious legal and privacy reasons) is very limited. It is the purpose of this paper however to analyse the feasibility of clustering notable artefacts rather than provide a definitive empirical study on how well this can be achieved more generally. In all four cases, the forensic images acquired are from the suspects' systems (not the victims'). This is important because the proposed novel algorithm in Section 5 is based upon an underlying assumption that a suspect will perform a series of criminally-related activities rather than a single action at any point in time (e.g. looking at child abuse imagery is likely to involve looking at many images rather than a single image in isolation or the composition of a letter for blackmail will result in the letter being emailed or printed rather than in being produced merely for the sake of it). On a victim's machine, this assumption is unlikely to hold true. The first public case (Case 1) was the "Hunter XP" provided by Guidance Software as a training case. This case is an artificial case describing a blackmail/stalking incident in which suspects demanded a ransom. The second public case (Case 2) was a hacking case that was artificially generated by National Institute of Standards and Technology (NIST) also for training purposes (NIST, 2013). Regarding the two private cases, they were obtained from the Sultanate of Oman - Public Prosecution. One was a document fraud case (Case 3) while the other one was an ATM skimming investigation (Case 4). A Non-Disclosure Agreement (NDA) was signed by the authors in order to maintain the confidentiality of the cases, thus preserving the privacy rights of those convicted criminals. These four cases were then manually analysed by an AccessData Certified Examiner (ACE) via FTK to provide a ground truth from which to compare and measure the performance of the approaches presented in this paper (i.e., which of the files were notable and which were noise (i.e., those are not relevant to the case)). The pre-processing options selected for each case included expanding compound files, data carving (on all pre-selected file types) across the complete image (allocated and unallocated), entropy test for encryption, and known file search (against the NSRL). Case details, including the image size and the total number of artefacts, are presented in Table 1.

Table 1. Case Details

Case ID	Image Size	Total Artefacts	Total notables	Proportion of notables (%)
1	500 MB	11,638	796	6.8
2	4.5 GB	22,373	11,696	52.2
3	16 GB	6,654	30	0.4
4	585 GB	3,456,219	281	0.008

During the analysis process, a list of files (referred to as the File List) was generated from the file system as they contain a rich source of information (e.g., MAC date and

time stamps, file path, or whether it was encrypted). Indeed, the File List contains the majority of files that are retrieved within the suspect's drive and their features, providing the most fundamental artefacts to digital evidence examiners. The File List was created after core forensic processes had already been undertaken. For examples, file hashing against the NSRL was performed and those that matched were removed (except for any alerted files). Data carving was also performed. However, the File List only presents the first level of metadata. It is recognised that many of the files themselves contain a rich source of metadata. It was deemed important to ensure clustering of these. Therefore, three application-level metadata files were also created based upon an analysis of their respective file repositories. All four metadata categories are presented in Table 2. Whilst this is not a definitive list of categories, with other application-level metadata categories possible such as Skype, Recycle Bin, Registry, and Office documents, these were deemed the most appropriate at this stage for the investigation. The features listed in Table 2 present the definitive features utilised in this paper. Please note, the encryption feature in the File List was determined by FTK based upon an entropy test being performed on each file and the duplicated flag based upon the result of the hash. Also the physical and logical sizes differ based upon the former being comprised of the logical plus any slack space to the end of the cluster.

Table 2. Features of different meta-data categories

Metadata Categories	Features
File List	Creation date and time, access date and time, modification date and time, file path, file extension, carved, deleted, encrypted, duplicated
Email	Subject, file name, to, from, cc, bcc, submit date and time, delivery date and time, unread, unsent, has attachment, physical size, logical size
EXIF	Last write date, last access date, date taken, camera make
Internet	Access date and time, file name, URL, number of hits

Upon the completion of the analysis process, information of the aforementioned four metadata categories were exported into individual Comma Separated Value (CSV) files for all cases. Each record within the CSV files was then marked as either notable or noise according to the result of the manual analysis process, providing a firm foundation for evaluating the performance of the proposed method. The marking of each entry was merely to establish how well SOM performed - this feature was *never* given to SOM or the subsequent process. Also, features of each record (as demonstrated in Table 2) were enumerated, allowing them to be processed by SOM. Due to the inability of SOM to process records with empty entries (e.g., files without a creation time), only records with timestamps were selected for a File List SOM – this meant carved files were excluded from the File List SOM. Whilst this might create the chance for files not being analysed by the File List SOM, it is envisaged that the application-level SOM analysis would include them. For example, in a case with a large number of carved JPEGs (such as Case 1), whilst the File List SOM would not include them, the EXIF metadata would

be included within an EXIF-based SOM. The total numbers of records that would be processed by the four separate SOMs are illustrated in Table 3. It is worth highlighting that due to the case nature, not all cases contain the four SOM categories. For instance, Case 3 only contains the File List information as the evidence was imaged from the suspect's USB stick; while Case 4 does not contain the email category as the machine was purposely built for committing the ATM skimming crime. The proportions of notables from the entire case that were processed by the SOMs are 95%, 8.6%, 100% and 100% for Cases 1 to 4 respectively (the figures were calculated based upon the information presented within Tables 1 and 3). Also, Case 2 was an interesting example of where a large number of notables were carved files (executable files related to hacking software), resulting them being excluded from the SOM analysis. The impact of this will be discussed in Section 6.

Table 3. The number of records for each category of the four cases (✓: Notable; ×: Noise)[1]

Case ID	File List SOM		Email SOM		EXIF SOM		Internet SOM		Total	
	✓	×	✓	×	✓	×	✓	×	✓	×
1	456	1,215	3	52	229	0	65	2,105	753	3,372
2	871	4,469	29	44	-	-	101	665	1,001	5,178
3	30	3,441	-	-	-	-	-	-	30	3,441
4	261	116,880	-	-	20	226	0	1,303	281	118,409

4 Experimental Methodology and Results for SOM Clustering

The purpose of the experiment was to determine whether metadata across a range of types is useful in automatically identifying notable versus noise files. Therefore, the experiment had two objectives:

- to investigate whether SOM can be utilised for clustering artefacts and if that were the case,
- to determine the influence of the network sizes upon the accuracy.

For each category (i.e., File List, Email, EXIF and Internet) within each of the four cases, the SOM neural network was configured with the following network sizes: 3x3 (9), 5x5 (25), 7x7 (49), 9x9 (81), and 10x10 (100). The SOMs were initialised randomly and with the aim of assuring the stability of the SOM result, the experiment was repeated 5 times for each metadata category of the four cases across all configurations. Hence, a total of 275 (11x5x5) sets of results were obtained. For each configuration, all records within a particular category (e.g., File List) were presented with all available features (as demonstrated in Tables 2 and 3). Details of these experimental results are presented and discussed in the following sections. The clusters were each analysed to identify the proportion of notables and noise files they contain.

[1] This also applies to Tables 4, 5, 6 and 7.

The experiment was conducted within the MATLAB R2013a environment on a Windows 7 Enterprise 64-bit Operating System with Intel Core i7-2600 CPU (3.4 GHz) and 16 GB memory (Matlab, 2015). MATLAB was chosen due to its ease in data manipulation and the availability of the SOM neural network.

4.1 File List

As illustrated in Table 4, in three of the four cases, clustering based upon the File List alone proved very successful. For cases 1, 3 and 4, all notable files were obtained within rank-five clusters with at least 56.4% noise files clusters in the remaining clusters by using the 3x3 SOM configuration. Indeed, Case 1 was able to identify 100% of notables (within three clusters) with 59.3% of the noise being grouped in the other six clusters. While Case 4, identified 100% of the notables in a single cluster and only introduced 1.6% of the noise. In comparison, only Case 2 did not identify all notable files within rank-five clusters – it was able to cluster 93.5% of notables at a cost of including 53.5% of the noise files – still resulting in a huge reduction in the number of files an investigator would need to analyse if these five clusters were successfully identified during an investigation. Also, the worst result was given by the 10x10 network in Case 2: only 32.4% of the notables were collected within the rank-five clusters, with the remaining 67.6% scattered around the other 95 clusters.

Table 4. Experimental results for the File List category of the 4 cases[2]

Network Size		9		25		49		81		100	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
1	1	88.8	15.3	38.4	5.4	21.9	1.6	10.3	0	11	1
	2	8.6	15.4	23.2	0.2	12.7	0	9.0	0.7	8.1	0
	3	2.6	10	18.2	1.6	11.2	1	8.6	0.3	7	0.1
	4	-	-	16.7	0	9.6	0	7.9	0	6.8	0
	5	-	-	2.6	9.3	9.4	0	7	0.3	5.7	0.2
	*****	**0**	**59.3**	**0.9**	**83.5**	**35.2**	**97.4**	**57.2**	**98.7**	**61.4**	**98.7**
2	1	38.7	18.9	16	5.3	13.1	2.1	9.9	0.9	9.9	0.9
	2	27.3	20.6	14.2	15	9.9	0.9	8.4	0.6	6	4.5
	3	14.8	2.9	12.4	1.9	8.5	10.9	7	0.1	5.7	0.1
	4	7.3	8.1	9.9	0.9	7.9	2.7	6.7	4.8	5.4	3
	5	5.4	3	7.9	3.3	7.3	3.3	5.4	3	5.4	1.7
	*****	**6.5**	**46.5**	**39.6**	**73.6**	**53.3**	**80.1**	**62.6**	**90.6**	**67.6**	**89.8**
3	1	56.7	8.7	56.7	8.8	43.3	3.3	43.3	3.3	23.3	1
	2	20	0.6	20	0.3	20	0.3	20	0.2	20	2.3

[2] ***:** Remaining Clusters; the best performances obtained by SOM are indicated by the grey shaded cells; the clusters are rank ordered based upon the proportion of notables grouped within them and the first 5 ranks are chosen for the demonstration purpose. This also applies to Tables 5, 6 and 7;

	3	10	6.6	10	6.6	10	3.8	13.3	2.2	20	0.3
	4	6.7	13.8	6.7	4.4	6.7	3.7	10	1.7	6.7	0.6
	5	6.7	13.9	6.7	4	6.7	4.6	6.7	1.6	6.7	1
	*****	**0**	**56.4**	**0**	**75.9**	**13.3**	**84.3**	**6.7**	**91**	**23.3**	**94.8**
4	1	100	16.3	100	3.8	100	2.3	100	1.9	100	1.6
	*****	**0**	**83.7**	**0**	**96.2**	**0**	**97.7**	**0**	**98.1**	**0**	**98.4**

Notably, as the SOM network size increases, both the identification rate of notable files and the volume of noise files decreases. For example, in Case 1, the percentage of notables from the 3x3 SOM to the 10x10 SOM reduces from 100% to 38.6% with a subsequent increase in the number of noise files being clustered from 59.3% to 98.7%. If this approach was used as a triage tool to identify whether notable files exist on the image (rather than identifying all notable files) this setting would provide 38.6% of notable files at a cost of 1.3% of noise files enabling the investigator to quickly understand the nature of evidence present.

4.2 Email

The results for the Email category where available are illustrated in Table 5. Please note, email was only present in two of the four cases. All the artefacts (both notables and noise) were grouped in one cluster for both cases; although around one quarter of the noise artefacts were separated from the notables for the network sizes 81 and 100 for Case 1. Reasons for this phenomenon could be due to the small amount of total email artefacts (53 and 73 for Cases 1 and 2 respectively), or a high level of similarities that were presented within them (i.e., the majority of them were clustered in a single cluster).

Table 5. Experimental results for the Email category of Cases 1 and 2

Network Size		9		25		49		81		100	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
1	1	100	100	100	100	100	100	100	76.9	100	73.1
	*	0	0	0	0	0	0	0	23.1	0	26.9
2	1	100	100	100	100	100	100	100	100	100	100
	*	0	0	0	0	0	0	0	0	0	0

4.3 EXIF

As illustrated by Table 6, all the EXIF files within Case 1 were notables; the results also highlight that SOM is capable of sorting data according to their similarities. In contrast with the results presented by Case 4, a better set of outcomes are observed as 90% of the notable files were grouped within two clusters by using the network size 9 configuration.

Table 6. Experimental results for the EXIF category of Cases 1 and 4

Network Size		9		25		49		81		100	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
1	1	49.3	-	15.7	-	10.0	-	7.4	-	9.2	-
	2	41.5	-	8.7	-	8.7	-	4.4	-	3.9	-
	3	2.6	-	7.9	-	7.4	-	4.4	-	3.5	-
	4	2.2	-	7.4	-	5.7	-	4.4	-	3.1	-
	5	2.2	-	7.4	-	5.7	-	3.5	-	3.1	-
	*	2.2	-	52.9	-	62.5	-	75.9	-	77.2	-
4	1	50	7.5	50	2.2	40	1.3	40	1.3	40	0
	2	40	3.5	40	1.3	35	0	35	0	30	1.3
	3	5	11.9	5	5.3	15	2.2	15	1.8	10	0
	4	5	11.9	5	6.2	5	0.4	5	0.4	10	2.2
	5	-	-	-	-	5	3.1	5	2.2	5	0.4
	*	0	65.2	0	85	0	93	0	94.3	5	96.1

Regarding Case 4, more than 95% of the notables can be found within the rank-five clusters for all network configurations; also the proportion of noise within the these five clusters reduces significantly as the network size increases: 34.8% of noise for the network size 9 in comparison with only 3.9% for the network size 100. Moreover, 100% of notable files can be observed under the network sizes 49, 81 and 100, reinforcing that SOM can be used for clustering information with a very high performance.

4.4 Internet

The results for Cases 1, 2 and 4 are presented in Table 7 (Case 3 did not contain any Internet activity). At least 78.5% and 55.4% of the notables were grouped within the rank-five clusters for Cases 1 and 2 respectively. The best performance (in terms of the proportion of notables) was obtained by using the network size 25 for Case 1: 100% of the notables were distributed in four clusters with only 18.8% of the total noise being clustered within. For Case 2, the worst performance was achieved under the configuration of the 10x10 network: only 55.4% of the notables were successfully clustered by the SOM within the rank-five clusters; however, due to the high density of notables within each cluster (4 with 100% of notables and 1 with 83.3% of notables), merely 2 noise artefacts (i.e., 0.3% of total noise in the case) were classified within those five clusters.

Table 7. Experimental results for the Internet category of Cases 1, 2 and 4

Network Size		9		25		49		81		100	
Case ID	Cluster ID	✓	×	✓	×	✓	×	✓	×	✓	×
1	1	35.4	13.5	33.8	3	33.8	2.5	23.1	1.1	23.1	1

	2	33.8	9.8	23.1	7	21.5	3.8	23.1	2.8	23.1	1.4
	3	20	3.9	23.1	4.9	20	3.9	13.8	1.9	12.3	1
	4	10.8	11.4	20	3.9	12.3	3.4	10.8	1.9	10.8	2
	5	-	-	-	-	10.8	3.8	9.2	0.9	9.2	0.3
	*	0	61.4	0	81.2	1.6	82.6	20	91.4	21.5	94.3
2	1	41.6	5.9	18.8	1.1	15.8	2.9	15.8	2.9	15.8	0
	2	25.7	7.5	17.8	0.2	14.9	0.2	10.9	0.2	11.9	0
	3	15.8	11	15.8	7.2	13.9	0.8	9.9	0.2	9.9	0.3
	4	14.9	11.3	14.9	1.7	12.9	0.6	9.9	0	8.9	0
	5	1	13.8	9.9	0.3	9.9	0.3	9.9	0.3	8.9	0
	*	1	50.5	22.8	89.5	32.6	95.2	43.6	96.4	44.6	99.7
4	1	-	15.7	-	12.7	-	4	-	3.7	-	3.5
	2	-	14.7	-	6.8	-	3.8	-	2.7	-	2.4
	3	-	14.3	-	5.7	-	3.5	-	2.5	-	2.3
	4	-	14.1	-	5.7	-	3.3	-	2.4	-	2.2
	5	-	14	-	5.7	-	3.2	-	2.2	-	1.8
	*	-	27.2	-	63.5	-	82.1	-	86.6	-	87.7

The Internet history of Case 4 was also processed despite no notables being presented within this category as in practice no one would know whether a file is notable or not unless it is examined. It helps to highlight the problem in practice of identifying the most appropriate network size – as a network size of 9 would result in an investigator having to analyse 72.8% of noise files, whereas a network size of 100 would offer the number of noise files to be analysed reduce to 12.3% if the rank-five clusters were selected.

4.5 Further Analysis

In general, a larger proportion of notables can be obtained with the rank-five clusters by using smaller SOM network sizes (e.g., 9 or 25); however, a considerable amount of noise files were also observed. A reduction in the noise can be achieved by choosing larger SOM network sizes (e.g., 81 or 100) with a compromise of the number of notables (as demonstrated in Fig. 1). Also, when the SOM network size increases, more clusters with a higher density of notables start to appear. Therefore, the granularity of the results is proportional to the sizes of the SOM network: smaller network sizes provide for coarser grained results while larger network sizes for finer grained outputs. This in part is obvious due to the larger number of clusters available. The need to be able to determine which network size to utilise is likely to be driven by the aim of the investigator (e.g., to obtain all notables at the cost of picking up more noise or obtain some notables to confirm the image has relevant content at the expense of very little noise).

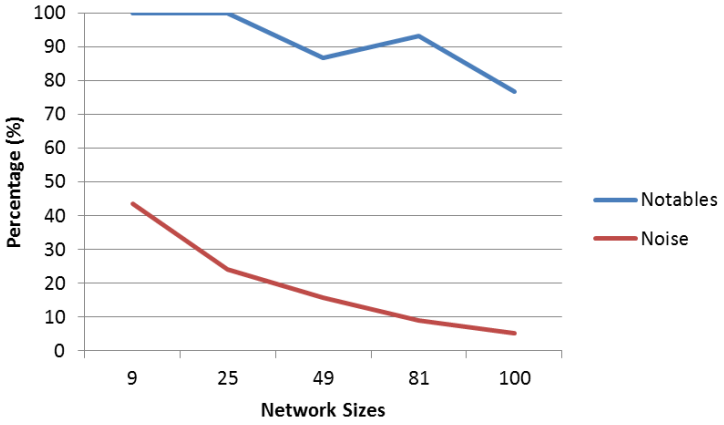

Fig. 1. Comparison on Notables and Noise in the rank-five clusters of the Case 3 File List on various network configurations

5 Automated Evidence Profiler

The previous section has highlighted that SOM-based analysis does have the potential to cluster notable versus noise files to a degree that would significantly reduce the investigation time. The research question that now arises is how to identify which clusters to analyse first – selecting the wrong clusters will lead the investigator to analyse a large number of noise files rather than notables. The following section describes a novel process for identifying and analysing relevant clusters and proceeds to present an evaluation of the approach.

5.1 Automated Evidence Profiler Process

The Automated Evidence Profiler (AEP) splits the problem of identifying relevant clusters into two aspects:

- How to identify the very first cluster to analyse
- How to identify subsequent clusters across the different metadata SOMs

Also, the AEP needs to maximise the number of notable files identified whilst looking to minimise the number of noise files included.

The solution to the first problem is based upon prior work completed in profiling criminal behaviour. It recognises that certain crime types will result in particular file formats being more likely to contain notable files than others. For example, child abuse cases will typically result in image-based files. The Department of Justice (DoJ) published a mapping of computer-based criminal activities versus file types (DoJ,

2001). The AEP uses this information to identify the first cluster within the File List SOM that contains the greatest number of these file types based upon the intelligence of the criminal activity.

In order to solve the issue of incorporating the results from the differing SOMs (File List and application-level SOMs) and to maximise the number of notable files identified, AEP proposes an approach that focuses upon analysing the timeline of files identified in the first cluster. The underlying assumption being applied is that when a suspect interacts with a file, the criminal activity is likely to involve other artefacts that are also notable within a short period of time (can be identified based upon their MAC times). For example, viewing a notable image is likely to involve viewing several images rather than a single image in isolation. Or the image is attached to an email and sent to another suspect. Fig. 2 illustrates the high-level process of AEP – from the initial crime mapping, to the creation of evidence trails and their subsequent prioritisation. The final step in the AEP process is to seek to improve upon the original crime mapping by feeding back into the process with the file types that have been identified as notable provide an adaptive and more reflective crime mapping. This process will also account for changes in criminal activities and how technology is used over time. For example, if a particular criminal activity begins to work with different file types, this process will account for this and ensure they are included. Whilst this feedback mechanism is proposed in this research, the evaluation did not include this aspect.

Fig. 2. Automated Evidence Profiler Process

The core of the AEP process is the creation of evidence trails and their prioritisation. As illustrated in Fig. 3, the initial identified cluster will give rise to a number of files

within it. Each file is taken in turn and a time window (e.g., 30 seconds, 1 minute, or 2 minutes) is applied to the accessed timestamp in both time directions (i.e., before and after) in order to create an evidence trail (i.e., finding any related artefacts within the given time window). Then these related artefacts are added to a global list of artefacts for the case. Having analysed all the files in the cluster, the resulting list of related artefacts is then sorted based upon the SOM in which the artefact resides and the cluster ID. The next cluster to be analysed is the one that appears most frequently within that global list. For example, as illustrated in Fig 3, based upon *Evidence Trail 1*, Internet Browser activity was identified as an action that took place shortly after the identified "Child.jpg" artefact (from the first cluster). If upon completion of the analysis of all the files from the first cluster, the cluster from the Internet SOM to which the *Facebook-Clare* exists appeared most frequently; that cluster would be the next one to be analysed. When it is, *Evidence Trail 4* will be created. The process then repeats as often as required.

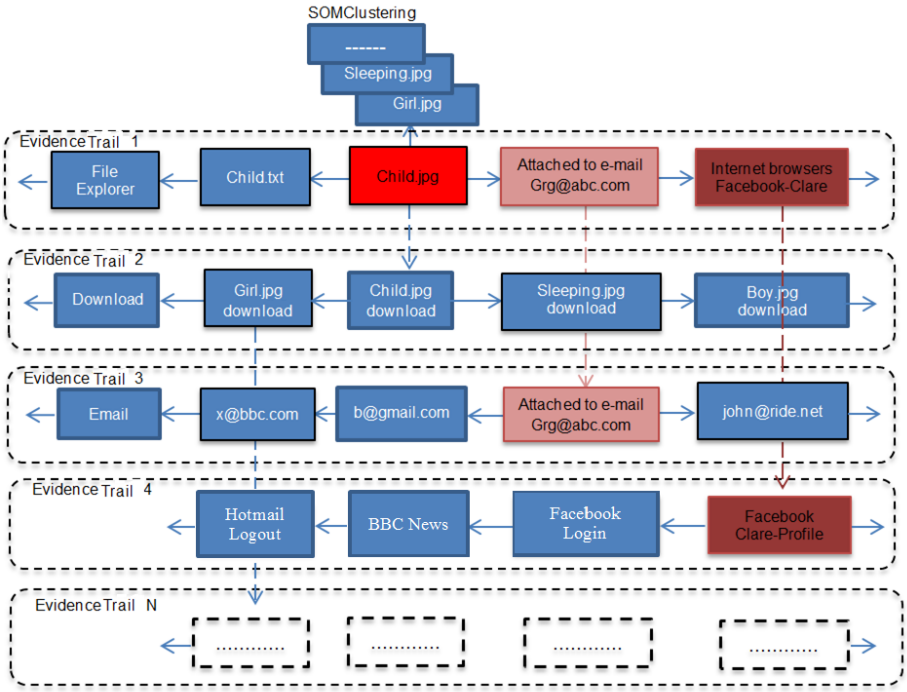

Fig. 3: Example of Evidence Trails

All of the analysed artefacts will be included in the final report to the investigator. The final aspect of the process is to provide some degree of prioritisation of the artefacts. This is achieved through an analysis of the frequency by which the artefact has been

identified – with reoccurring artefacts more likely to be key notable artefacts that a suspect has interacted with in comparison to infrequent used artefacts.

5.2 AEP Evaluation

An evaluation of the AEP process was performed utilising the aforementioned cases presented in section 3. Within the evaluation, impacts of three key parameters of the AEP are also tested:

- Time Window (TW) – whilst the assumption that criminal activities are not undertaken in isolation may intuitively seem correct, the size of the TW will have an impact on the number of notables and noise files (with the former decreasing and the latter increasing if the TW is too large).
- Number of clusters to be analysed – how many iterations of the process are necessary before the number of notable artefacts diminishes and the number of noise files increases. Too few iterations will result in insufficient notable artefacts being identified – too many and the process will present more noise files for the approach to be useful.
- Network size – as indicated in the previous section, the relationship between notables and noise files are closely related to the network size

In addition, it is also important to investigate whether the AEP process is able to incorporate the results of the other SOMs – rather than merely analysing clusters within the File List SOM.

Table 8 illustrates the best-case results (and the algorithmic parameters required for the result). Given the only investigator interaction required in this process is to select the crime category, the process achieved recall rates (proportion of notable files identified) of between 75.3 to 93%. As the main aim of the digital investigation is to find as much evidence as possible, the recall rate achieved by the AEP is promising as at least three quarters of the notables files were identified across the four cases. While the precision rate is low in terms of what an average clustering technique enabled system can achieve, this phenomena could be caused by a number of factors, including the use of smaller network sizes, the lower number of notable artefacts, and the proportion of notables within the case (e.g. only 0.008% of the artefacts were notables in Case 4). It is worth highlighting, in terms of the computational time required, most of the analysis took a relatively trivial amount of time (in the order of minutes); however, Case 4 (with a large number of artefacts to process) did take on average an hour to process. In comparison to many forensic analyses this is not significant; however, it is worth highlighting and noting as an opportunity for future research. Overall, the investigator can still save huge amount of time when the AEP process is applied to analyse the case.

Table 8. Overall Performance of the AEP Process

Case	Recall	Precision	F	Process Parameters

ID	(%)	(%)	measure	SOM Network Size	TW (minute)	SOM Iteration	Time (minute)
1	93	34.4	0.5	9	2	3	1
2	75.3	36.3	0.49	9	0.5	2	7
3	76.7	12.9	0.22	49	2	2	1
4	92.8	0.5	0.01	100	0.5	1	~60

Fig. 4 illustrates the impact upon performance the TW has across differing network sizes. Generally as expected, beyond a certain TW, the proportion of noise files significantly increases with a disproportionate change in the notable artefacts identified. There appears to be a "sweet spot" at around the 2-5 minute window.

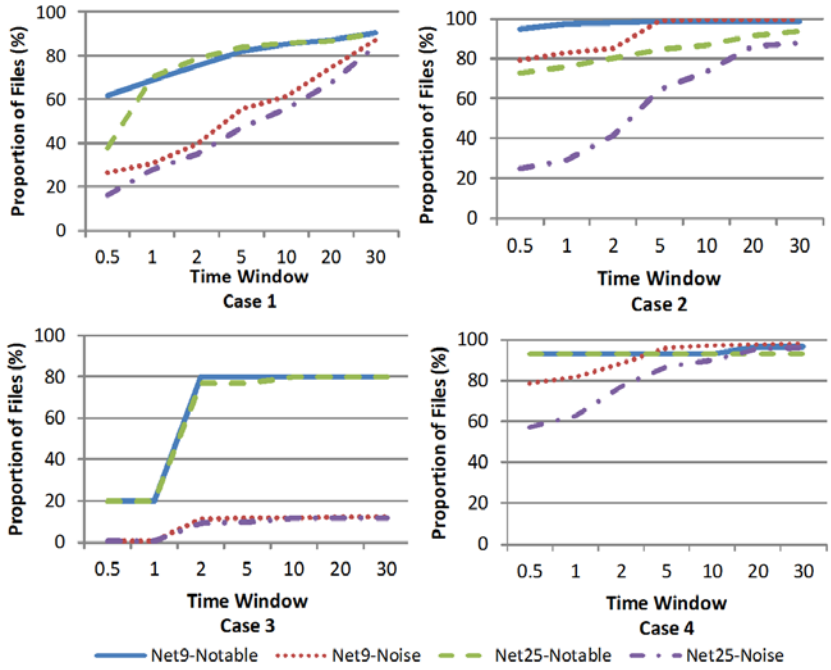

Fig. 4: Impact of the Time Window on Performance

In terms of how many SOM iterations should be analysed, the graphs illustrated in Fig. 5 show it is more dependent on the size of the SOM network utilised, which in turn appears to have a dependency on the number of artefacts to be processed. Cases with larger numbers of artefacts appear to operate better in larger SOM network configurations (although the limited number of cases analysed merely suggest this rather than definitively indicate it).

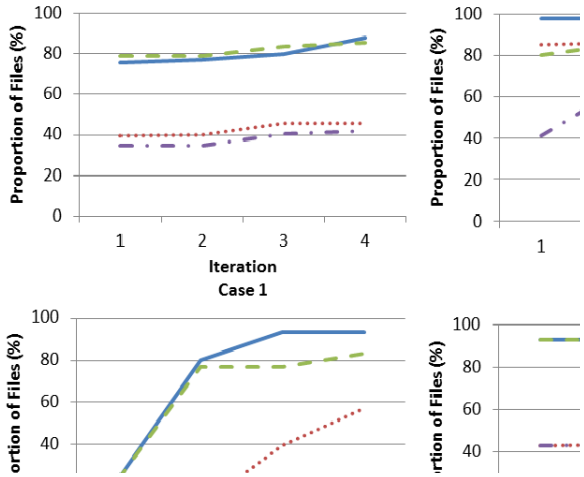

Fig. 5. Impact of Cluster Iterations on Performance

The final aspect of the investigation was to ensure the process was able to incorporate notable files that are contained in clusters pertaining to the other SOMs (i.e., EXIF, Internet and Email). For simplicity, Table 9 illustrates the effect of the process upon Case 1 and 2 versus three network configurations with four iterations. The first number indicates the SOM type and the second the cluster ID. In both of these cases it is clear that clusters from all relevant SOMs (where they exist) are being included within the analysis for notable files. For example, clusters from all four SOMs are included in Case 1.

Table 9. SOM and Cluster IDs Identified

Network Size		9				25				49			
Case ID	TW	1[3]	2	3	4	1	2	3	4	1	2	3	4
1	0.5	1,7	1,5	4,3	1,8	1,1	1,4	1,5	1,20	1,18	1,21	3,1	1,42
	1	1,7	1,5	4,3	1,8	1,1	1,25	3,17	3,21	1,18	1,6	3,1	3,8
	2	1,1	1,2	1,6	3,4	1,1	2,17	1,7	1,3	1,25	1,1	1,2	1,27
	5	1,7	1,8	1,6	1,3	1,1	2,13	1,15	3,1	1,1	1,4	1,19	4,42
	10	1,7	4,5	1,8	4,3	1,3	3,1	3,6	3,10	1,33	2,1	2,4	4,23
	20	1,7	4,5	1,8	1,6	1,2	3,15	3,16	3,22	1,18	4,46	4,24	4,17

[3] Iteration ID

	30	1,1	4,5	1,8	1,6	1,1	2,25	3,2	3,3	1,31	3,3	3,4	3,10
2	0.5	1,1	1,4	1,7	1,6	1,7	1,10	1,16	1,1	1,22	1,47	1,43	1,12
	1	1,1	1,4	1,7	1,5	1,7	1,16	1,25	1,13	1,22	1,39	1,47	1,48
	2	1,4	1,1	1,5	1,7	1,7	1,25	1,16	1,21	1,22	1,47	1,40	1,48
	5	1,1	1,4	1,5	1,7	1,7	1,13	1,16	1,25	1,22	1,36	1,27	1,8
	10	1,1	1,4	1,5	1,7	1,7	1,16	1,25	1,2	1,22	1,43	1,37	1,27
	20	1,1	1,4	2,9	4,5	1,7	4,14	4,19	4,18	1,22	1,43	4,20	4,13
	30	1,1	4,3	4,5	2,9	1,7	4,14	4,19	4,18	1,22	1,43	4,7	4,13

(Key: 1=File List SOM; 2=Email SOM; 3=EXIF SOM; 4=Internet SOM)

6 Discussion

Based upon the results presented in Tables 4-7, the application of SOM to artefact identification works well – showing that notable artefacts can be correlated based upon metadata that is derived from it. Indeed, for each of the categories (i.e., File List, Email, EXIF, and Internet) through all four cases, more than 93.5% of notables were grouped within the rank-five clusters at least under one SOM network configuration with at less than half of the noise files being included. The best performance of all results (in terms of grouping most notables within the rank-five clusters and also minimising the proportion of noise) was achieved by using the 10x10 network for the File List in Case 4 – all the notables were clustered in a single cell with only 1.6% of the total noise being present.

Case 2 has proven to be challenging, both in the identification of notable versus noise files but also in the inclusion of sufficient notable files in the first place. The latter problem has the potential to be rectified through the inclusion of more application-level metadata SOMs. For example, a Recycle Bin SOM that extracts the metadata from the INFO2 (pre Windows Vista) or $I (Windows Vista and newer) would have identified previously deleted applications (as was the issue in Case 2). The same argument could be made for a variety of other application-level SOMs, including Microsoft Office files where metadata is extracted from the header, Skype SOM that extracts Skype call records and chat interactions, and Log SOMs that extract application and user based log records. The reason for the poor identification performance with the artefacts that were included in Case 2 is less clear; however, it is notable that having applied the AEP process the overall performance was commensurate with the remaining cases – achieving a recall rate of 75.3%. So whilst the SOM might have had some weaknesses, the proposed AEP process has counteracted these.

The results from the AEP more generally are also very encouraging. Whilst ideally it would best to have identified all notable files (i.e., 100% recall rate), an appreciation of the issues regarding the volume of data, time taken to investigate, and the likelihood of human investigative error means the opportunity of undertaking a “gold standard” investigation is becoming less and less likely in the future. Therefore, more “intelligent” approaches, such as those proposed in this paper, provide a basis upon which simpler more technically trivial cases (which represent a large proportion of the day-to-day activities for law enforcement forensic investigators) can be undertaken, reducing (not

removing or replacing) the time taken for an investigator to confirm the nature of the case and to present the relevant evidence in a report.

7 Conclusions and Future Work

The paper has presented an empirical study investigating the possibility of using SOM to cluster notable files for digital forensic investigations in an unsupervised manner – a significant enhancement over existing approaches. The experimental results show the analysis of file system and application-level metadata offers a good level of performance. It should be highlighted that in contrary to much of the literature in digital forensics, the purpose of this paper is to further the body of knowledge in the application of machine learning to digital forensics for the purpose of automated artefact identification. As such the paper has presented an investigation in the approach, developed a model and evaluated the model parameters to illustrate the impact upon performance. It is not the objective of the paper to present a forensic tool, nor to advocate that such an approach would remove the necessity of an investigator. The approach presented is the building block upon which further research can seek to refine and evaluate with a view to providing a triage tool that could assist in a range of computer-related crimes. The net effect will be to free up more investigative time for more technically complex cases (such as those involving the use of advanced data hiding and anti-forensics).

Future work will focus upon investigating the effectiveness of other unsupervised machine learning approaches and seek to evaluate the proposed AEP process against a greater range of cases. Further work also needs to be undertaken to investigate the wider architectural aspects of implementing such an approach in practice – for example, the visualisation of the identified artefacts in order to enable a faster appreciation of their status and relevance by an investigator. This could also include a feedback loop that allows the system to adapt to changes in the identification process given by decisions from a trained and experienced investigator.

References

AccessData (2015) **Forensics Toolkit**, http://accessdata.com/solutions/digital-forensics/forensic-toolkit-ftk

Adderley, R., and Musgrove P.B. (2001) **Data mining case study: modeling the behavior of offenders who commit serious sexual assaults**, Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, August 26-29, San Francisco, CA, USA. ACM, 2001, pp. 215-220, 2001

Beebe, N.L. and Clark, J.G. (2007) **Digital forensic text string searching: Improving information retrieval effectiveness by thematically clustering search results**, Digital Investigation 4, no. Supplement 1, 49–54.

Beebe, N.L, Clark, J.G., Dietrich, G.B., Ko, M.S. and Ko, D. (2011) **Post-retrieval search hit clustering to improve information retrieval effectiveness: Two digital forensics case studies**, Decision Support Systems, volume 51, Issue 4, pages 732-744

Brainz (2014) **15 Criminal Cases Solved With Digital Evidence**, http://brainz.org/15-criminal-cases-solved-digital-evidence/

Carrier, B (2015) **Autopsy**, http://www.sleuthkit.org/autopsy

Casey, E. ed (2010) **Handbook of Digital Forensics and Investigation**, Academic Press. p. 567. ISBN 0-12-374267-6

Casey E., Stellatos G.J. (2008) **The impact of full disk encryption on digital forensics**, ACM SIGOPS Operating Systems Review, Vol42 Issue (3). Page 93-98 DOI: 10.1145/1368506.1368519

DoJ (2001) **Electronic Crime Scene Investigation: An On-the-Scene Reference for First Responders**, US Department of Justice, https://www.ncjrs.gov/pdffiles1/nij/227050.pdf

e-fense (2014) **products**, http://www.e-fense.com/products.php

FBI (2013) **Regional Computer Forensics Laboratory Annual Report for Fiscal Year 2013**, https://www.rcfl.gov/downloads/documents/fiscal-year-2013/view

Fei, B., Eloff, J., Olivier, M. and Venter, H. (2006) **The use of self-organising maps for anomalous behaviour detection in a digital investigation**, Forensic Science International 162, no. 1-3, 33–37.

Fei, B., Eloff, J., Venter, H., and Olivier, M. (2005) **Exploring forensic data with selforganizing maps, Advances in Digital Forensics**, Springer, pp. 113–123.

Feyereisl, J. and Aickelin, U. (2009) **Self-Organising Maps in Computer Security**, In Computer Security: Intrusion, Detection and Prevention, Ed. Ronald D. Hopkins, Wesley P. Tokere, pp. 1-30, Nova Science Publishers.

Garfinkel, S. (2010) **Digital forensics research: The next 10 years**, Digital Investigation 7, S64-S73

Guidance Software (2015) **EnCase Forensic v7**, https://www.guidancesoftware.com/products/Pages/encase-forensic/overview.aspx

HHD Software (2015) **Free Hex Editor Neo**, http://www.hhdsoftware.com/free-hex-editor

Inforsecusa (2011) **Computer Forensics Criminal Cases**, http://infosecusa.com/computer-forensics-criminal-cases

Kohonen, T. (1998) **The self-organizing map**, Neurocomputing, Vol. 21, Issues 1-3, pages 1-6

Kayacik, H.G, and Zincir-Heywood, A.N. (2006) **Using self-organizing maps to build an attack map for forensic analysis**, proceeding of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services, article no. 33 doi: 10.1145/1501434.1501474

Lawton, D., Stacey, R., Dodd, G. (2014) **eDiscovery in digital forensic investigations**, technical report 32 UK Home Office, London, URL: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/394779/ediscovery-digital-forensic-investigations-3214.pdf

Matlab (2015) **R2013b Release Highlights**, http://uk.mathworks.com/products/new_products/release2013b.html

NIST (2013) **The CFReDS Project**, http://www.cfreds.nist.gov/

Office of the Inspector General (2015) **Audit of the Federal Bureau of Investigation's Philadelphia Regional Computer Forensic Laboratory, Radnor, Pennsylvania, Audit Report**, https://oig.justice.gov/reports/2015/a1514.pdf

Palomo, E.J., North, J., Elizondo, D., Luque, R.M. and Watson, T. (2011) **Visualization of network forensics traffic data with self-organizing map for qualitative features**, Proceedings of the International Joint Conference on Neural Networks, IEEE, San Jose, California, USA, pp. 1740-1747

Paraben Corporation (2015) **Computer Forensics/ P2C**, https://www.paraben.com/p2-commander.html

RCFL (2014) **News**, https://www.rcfl.gov/news

Veracode (2015) **Business and Economic Consequences of Inadequate Cyber-security**, Veracode and Centre for Economics and Business Research Ltd, https://info.veracode.co.uk/analyst-report-cebr-business-and-economic-consequences-of-inadequate-cybersecurity.html

Wang, W.B., Huang, M.L., Zhang, J., and Lai, W. (2015) **Detecting Criminal Relationships through SOM Visual Analytics**, inInformation Visualisation (iV), 2015 19th International Conference on , vol., no., pp.316-321, 22-24 July 2015 doi: 10.1109/iV.2015.62