Digital Investigation 18 (2016) 65–76

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

A suspect-oriented intelligent and automated computer forensic analysis

M. Al Fahdi [a], N.L. Clarke [a, b], F. Li [a, *], S.M. Furnell [a, b, c]

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

ARTICLE INFO

Article history:
Received 22 March 2016
Received in revised form 19 July 2016
Accepted 15 August 2016
Available online 16 August 2016

Keywords:
SOM
Digital forensics
Automation
Clustering
Self-organising map
Cybercrime

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 10×10 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%.

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

* Corresponding author.
E-mail address: info@cscan.org (F. Li).

http://dx.doi.org/10.1016/j.diin.2016.08.001
1742-2876/

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 use of cloud computing (Casey and Stellatos, 2008; Garfinkel, 2010). Statistics from the FBI's Regional Computer Forensics Laboratory (RCFL) show that they had processed 5973 TBs of data from 7273 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 1566 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 "Related work" presents the existing work surrounding the use of SOM clustering with respect to digital forensics. Section "Datasets" describes the datasets that were utilised in the experiment, with Section "Experimental methodology and results for SOM clustering" presenting the experimental results of the SOM study. Section "Automated Evidence Profiler" 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 "Discussion", prior to the conclusion and future work.

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, 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 2640 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