Towards fully automated forensic file identification with approximate matching algorithms

Elsevier[1]

Radarweg 29, Amsterdam

Elsevier Inc[a,b], Global Customer Service[b,]*

[a]1600 John F Kennedy Boulevard, Philadelphia
[b]360 Park Avenue South, New York

Abstract

The automated analysis of large amounts of data is a highly researched field in modern digital forensics. Against the backdrop of the ever-increasing mass of data to be processed, it is necessary to find ways of evaluating the data automatically. In scenarios where the loss of sensitive data is being investigated, procedures that help to identify files among many others are precious. The genre of approximate matching algorithms that address this challenge has seen much innovation in recent years. The central question: "How can I determine the presence of a target file on a disk image?" has found an answer in several algorithms. This paper will now examine for the first time their practicality in an automated evaluation pipeline. The aim is not only to reduce the target set of digital artifacts to the most promising ones and then to evaluate them manually as it happens in triage, but to automatically identify one particular one among many. Our research shows which approximate matching algorithm is best suited for identifying which file type. Furthermore, robustness and efficiency are assessed. The result of our research is a tool called ApproxIdentifier that takes a target file and a hard disk image as input and can detect the presence of the file on the disk image efficiently and without further input. It benefits not only from the research on approximate matching but also from the latest insights on file carvers and bulk extraction. As a test, synthesized hard disk images with target files on them are generated. Using file type identification and other means of bulk extraction, the target set of files is narrowed down and then, depending on the data type, passed to the most appropriate approximate matching algorithm, which performs the unambiguous identification. Our research is intended to assist forensic scientists in the targeted search for digital artifacts like malware traces and sensitive documents.

Keywords: `elsarticle.cls`, LaTeX, Elsevier, template
2010 MSC: 00-01, 99-00

1. Introduction

Digital forensics has become more difficult due to the capacity and diversity of different devices that contain digital evidence. Many approaches have been proposed to cope with the onslaught of data, including: parallelization and multi-processing as well as statistical sampling or even extending the time a suspect can be detained without charge so that evidence can be analyzed. Regardless of these methods, however, it is clear that as long as a backlog exists, a process must be created to allocate limited forensic resources. Normally, an investigator follows a suspicion that has been formed by other evidence. Mosty there is no concrete piece of evidence or a specific digital artifact known prior to the investigation that has to be found. Sometimes, however, it is necessary to search for a specific piece of evidence, for example in the case of securing evidence after a malware attack. The type of malware may be known, including the payload but now it is necessary to search for the document that served as a initial vector onto the digital premises. Another example are the investigations in the cases of unauthorized possession of documents or their disclosure through a leaker. An investigator may be in possession of reference documents whos presence on a hard drive would conclusively prove a unauthorized posession. These kinds of investigations that aim to find a concrete digital artifact might be fewer than those where an investigator starts on a level playing field and follows up on a suspicion. However, the procedure in the initial phase is the same in both cases. First, the basic set of files can be differentiated with the help of

[★] Fully documented templates are available in the elsarticle package on CTAN.

[*] Corresponding author

Email address: `support@elsevier.com` (Global Customer Service)

URL: `www.elsevier.com` (Elsevier Inc)

[1] Since 1880.