Digital Investigation 11 (2014) S18–S26

Contents lists available at ScienceDirect

Digital Investigation

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

Practical use of Approximate Hash Based Matching in digital investigations

Petter Christian Bjelland*, Katrin Franke, André Årnes [1]

Norwegian Information Security Laboratory (NISlab), Gjøvik University College, Norway

Keywords:
Digital forensics
Approximate Matching
Evidence analysis
Data discovery
Malware forensics

A B S T R A C T

Approximate Hash Based Matching (AHBM), also known as *Fuzzy Hashing*, is used to identify complex and unstructured data that has a certain amount of byte-level similarity. Common use cases include the identification of updated versions of documents and fragments recovered from memory or deleted files. Though several algorithms exist, there has not yet been an extensive focus on its practical use in digital investigations. The paper addresses the research question: *How can AHBM be applied in digital investigations?* It focuses on common scenarios in which AHBM can be applied, as well as the potential significance of its results. First, an assessment of AHBM for digital investigations with respect to existing algorithms and requirements for efficiency and precision is given. Then follows a description of scenarios in which it can be applied. The paper presents three modes of operation for Approximate Matching, namely *searching*, *streaming* and *clustering*. Each of the modes are tested in practical experiments. The results show that AHBM has great potential for helping investigators discover information based on data similarity. Three open source tools were implemented during the research leading up to this paper: *Autopsy AHBM* enables AHBM in an existing digital investigation framework, *sddiff* helps understanding AHBM results through visualization, and *makecluster* improves analysis of graphs generated from large datasets by storing each disjunct cluster separately.

Introduction

The focus of this study has been to assist investigators in law enforcement to organize and analyze digital evidence using *Approximate Matching*. The term Approximate Matching refers to the technique of detecting data that are in some way similar. Though tools for performing Approximate Matching of raw data have been known for some time, they are still not integrated in popular digital investigation tools. Approximate Matching has been a standalone capability only used under special circumstances and not as part of standard investigation practices. Why AHBM is not yet in widespread use is difficult to determine, however, two reasons may be dominant: *No integration with existing digital investigation tools*, and *limited knowledge of the potential gains of using it*.

There are three types of Approximate Matching: perceptual, content and hash based matching. While the two latter focus on identifying data that is similar from the perspective of a computer, perceptual matching identifies data that is similar from the perspective of a human. Whereas perceptual matching is well suited for comparing pictures and videos, content and hash based matching algorithms are designed to match binary data, such as documents, executables, memory dumps and network traffic. Hash based matching groups chunks of data and compare them with chunks in other files. Content based matching

* Corresponding author.
E-mail addresses: petter.bjelland@hig.no (P.C. Bjelland), katrin.franke@hig.no (K. Franke), andre.arnes@hig.no (A. Årnes).
[1] The author A. Årnes is also associated with Telenor Group.

http://dx.doi.org/10.1016/j.diin.2014.03.003

computes the exact difference between two files, often using techniques such as Hamming distance (Hamming, 1950) and Levenshtein distance (Levenshtein, 1966). All these types of Approximate Matching may be relevant for digital investigations.

Related to this study is the work by Vassil Roussev and Candice Quates on hash based matching using empirical models for identifying, representing and matching chunks of data with their tool *sdhash* (Roussev, 2010, 2011). They also present an evaluation and comparison of existing AHBM tools in Roussev (2011). Most other published work on AHBM techniques focuses on the technical aspects of how hash based similarity can be measured. An exception is the forensic investigation of the M57 dataset (DigitalCorpora.org, 2009) by Roussev and Quates (2012). The authors describe how *sdhash* can be used to analyze large amounts of data, with a particular focus on reducing the amount of data subjected to human analysis. They present three scenarios in which Approximate Matching can be applied: Detecting the presence of contraband, detecting unauthorized copying of internal data, and detecting unauthorized exfiltration of data. The study focuses on evidence detection and what questions an analyst may ask using AHBM techniques. In contrast, this study focuses less on the inner workings of any particular tool, and instead attempts to define a general modus operandi for Approximate Matching when applied to digital investigations.

This research also has resemblance to cross-evidence correlation techniques different from AHBM, such as large scale data triage (Garfinkel, 2013) and malware identification (Flaglien et al., 2011).

For the purpose of the experiments in this paper, *sdhash* was used as the approximate hash based matching tool, rather than *ssdeep*. This is because *sdhash* yields the most robust and accurate results, as shown in Breitinger et al. (2013) and Roussev (2011). However, as *ssdeep* is more efficient when matching files without specialized hardware (Breitinger et al., 2013) (*sdhash* comparison efficiency depends on whether the POPCNT CPU instruction is available or not), it may be the preferred tool in many situations.

A system for comparing AHBM algorithms, named FRASH[2] was proposed by Breitinger et al. (2013). This work is important for the introduction of Approximate Matching as an integrated part of digital investigations, but current efforts have been limited to reviewing the types of similarities that may be discovered. In order for these techniques to become widely applied in digital forensics, there is a need to explore when and how to approximately match available evidence to discover new information.

This paper addresses these issues by describing the common scenarios where AHBM may add unique information to an investigation. Through practical experiments, the relevance of using the tools in the defined scenarios are reviewed and discussed.

During the research leading up to this paper, three tools were implemented and made available in order to help performing AHBM and analyzing its results. First, a module for performing AHBM in the digital investigation framework Autopsy (Carrier, 2012) was implemented.[3] The module, called *Autopsy AHBM*, allows an investigator to easily perform AHBM searching and streaming during disk image analysis. Second, a tool called *sddiff*[4] was implemented to help understand similarity through visualization. Finally, *makecluster*[5] was implemented to enable analysis of individual clusters by storing each connected cluster in separate files. The tool makes it easier to analyze graphs generated from large datasets, both in terms of computational complexity and visualization.

Fig. 1. Two semantically and perceptually similar files. The two files are not syntactically similar. Color to the left and grayscale to the right. (Photography captured by Petter Chr. Bjelland, 2013).

In the following, Section What is similarity? addresses the philosophical question: What is similarity? Section Approximate Matching describes in detail the various types of Approximate Matching. Section Modes of Approximate Hash Based Matching describes the different modes in which AHBM can be used, and what knowledge and insight may be achieved using modes. Then, Section Practical scenarios with AHBM describes scenarios for what types of data may be analyzed using these modes. Finally, the last two sections complete the paper with subjects for further research and conclusions.

What is similarity?

A word frequently used when discussing Approximate Matching is *similarity*. Investigators may use Approximate Matching to discover the presence of files *similar* to something we already know. So what is similarity and how do we measure it? There are essentially two different ways in which two files can be *similar*: syntactic and semantic. Syntactic similarity is from the perspective of a computer, and semantic similarity is from the perspective of a human.

Two documents are semantically identical if they communicate the same information. For example, a Microsoft PowerPoint presentation is semantically identical to an exported PDF document containing the same pages. Their cryptographic hashes[6] will not be identical, though we can still argue that the documents are the same. A similar concept applies to media, like pictures and videos.

[2] FRASH: A framework to test algorithms of similarity hashing.

[3] https://github.com/pcbje/autopsy-ahbm.

[4] https://github.com/pcbje/sddiff.

[5] https://github.com/pcbje/makecluster.

[6] A cryptographic hash is a digital fingerprint, making it possible to determine whether to pieces of data are exactly the same.