Digital Investigation 24 (2018) S121–S130

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2018 Europe — Proceedings of the Fifth Annual DFRWS Europe

A standardized corpus for SQLite database forensics

Sebastian Nemetz, Sven Schmitt*, Felix Freiling

Computer Science Department, Friedrich-Alexander-University Erlangen-Nuremberg (FAU), Germany

Keywords:
Digital forensics
Database forensics
Forensic corpora
SQLite3

ABSTRACT

An increasing number of programs like browsers or smartphone apps are using SQLite3 databases to store application data. In many cases, such data is of high value during a forensic investigation. Therefore, various tools have been developed that claim to support rigorous forensic analysis of SQLite database files, claims that are not supported by appropriate evidence. We present a standardized corpus of SQLite files that can be used to evaluate and benchmark analysis methods and tools. The corpus contains databases which use special features of the SQLite file format or contain potential pitfalls to detect errors in forensic programs. We apply our corpus to a set of six available tools and evaluate their strengths and weaknesses. In particular, we show that none of these tools can reliably handle all corner cases of the SQLite3 format.

Introduction

SQLite has evolved into one of the most widely used database management systems (DBMS) in the world (SQLite Online Documentation, 2016a). Originally targeting the domain of embedded devices, its standalone DBMS implementation with comprehensive support for SQL made it a very popular storage engine for messaging applications and browsers most prominently (but not only) on mobile devices. The widespread acceptance of SQLite has lead to the situation where digital investigations regularly require the forensic analysis of data stored in SQLite databases.

Motivation

As in many other areas of forensic computing, the first tools for the forensic analysis of SQLite files were created by practicioners in response to urgent needs. Commercial tool vendors contributed their own share of responses to imminent market needs. Little, however, is known about the level of reliability and scrutiny with which these tools are able to analyze general SQLite files. A detailed knowledge of the strengths and weaknesses of tools is, however, required within a forensic analysis.

On the importance of forensic corpora

With respect to forensic tool development and tool testing, Garfinkel et al. (2009) argued that digital forensic science should strive for reproducibility of tool evaluations by adopting performance requirements and standards for forensic software. Basic concepts to realize this claim are the development and use of standardized corpora, which have been influential and beneficial in many other areas of computer security such as intrusion detection systems (Lippmann et al., 2000).

Without standardized corpora, contributions often rely on self-collected, personal and rather small sets of data, sometimes only created for a very particular purpose. Such data sets are typically not available to the public. Other researchers can neither reproduce, verify, nor build upon the test cases and results presented. There is no way to objectively compare the strengths and weaknesses of different tools with each other, without a test data set that is standardized and publicly available. In short, research efforts and results that are performed on individual data sets are not comparable and therefore less useful. Such efforts have limited impact and are basically lost for future research.

A more scientific approach is to make data sets available to the public. This way, they can be used by different people for different purposes over time. Exemplary purposes are testing and evaluating existing tools or even developing new ones. With a publicly available test dataset, new approaches and algorithms for forensic analysis can directly be compared to existing tools. Possible improvements, such as less ressource consumption, improved performance or better analysis results, can be measured and made available to everyone. The acceptance of a tool by its users has no longer to be based on the reputation of its creator(s), but can be derived from publicly available and reproducible test results. The users can have increased confidence in the tool and review its quality. This way, public review of software (tools) can make them

* Corresponding author.
E-mail addresses: sebastian.nemetz@posteo.de (S. Nemetz), sven.schmitt@cs.fau.de (S. Schmitt), felix.freiling@cs.fau.de (F. Freiling).

https://doi.org/10.1016/j.diin.2018.01.015

more robust, more reliable and more trustworthy. This is what we should request and claim for tools, especially when used in forensic investigations. We give an overview about existing corpora in the following section.

Related work

Many standardized corpora in digital forensics exist today (Yannikos et al., 2014; Digital Corpora, 2009) and are publicly available for download. Starting with a study of disk sanitization practices, Garfinkel (Garfinkel and Abhi, 2003) published a sequence of papers relating to the importance and benefits of standardized corpora in forensic computing (Garfinkel, 2006, 2007; Garfinkel et al., 2009; Zarate et al., 2014). In 1998, Garfinkel started to buy and collect hard drives from the second hand markets and thereby established the *Real Data Corpus* that, today, comprises more than 35 TB of data. This corpus aims at providing general hard disk images and a high variety of file types to test digital repository architectures (Woods et al., 2011), general disk forensics tools (Guo et al., 2009; Garfinkel, 2012a) and automation processes (Garfinkel, 2009, 2012b). The Real Data Corpus was, however, not designed to specifically test SQLite analysis tools. A comprehensive overview of available datasets with relevance to digital forensics is presented by Grajeda et al. (2017)

Contributions

In this paper, we introduce the – to our knowledge – first forensic corpus specific to the SQLite DBMS. With *SQLite* we actually refer to SQLite Version 3 (SQLite Online Documentation, 2004), which was released in 2004 and is still the most recent and widely spread version. More specifically, we make the following contributions:

- We dissect the SQLite3 file format and describe characteristics that are important with regard to forensic analysis and analysis tool robustness. While some may appear exceptional, all cases do actually fully comply with the definition of the official file format (SQLite Online Documentation, 2016b).
- Based on the file format analysis, we develop a forensic corpus of SQLite database files focussing on the inner structures of the database file format. The corpus consists of 77 databases grouped into 5 categories according to their peculiarities. Along with the database files, the corpus also comprises a technical documentation about the creation of every single database and its contents (ground truth).
- We apply the newly created corpus to a set of forensic software able to process SQLite3 database files. We thereby evaluate strengths and weaknesses of existing tools and show that none of the available tools can handle all corner cases reliably.

We make this corpus accessible for further research by making it available online at the following URL: https://faui1-files.cs.fau.de/public/sqlite-forensic-corpus/.

Outline of this paper

The paper is structured as follows: First, we give an overview about general characteristics and categorization of corpora (Section: Background on forensic corpora). Subsequently, we introduce the SQLite Forensic Corpus (Section: Introducing the SQLite Forensic Corpus) and present its structure as well as accompanying metadata. Details about the individual contents and peculiarities for each of the 77 database files comprised in the corpus are then provided (Section: Databases in the SQLite Forensic Corpus). These define different scenarios against which any SQLite processing tool can be tested. We evaluate several SQLite specific tools against the newly introduced corpus and discuss the test results (Section: Evaluation). Finally, we summarize the paper and conclude (Section: Conclusion).

Background on forensic corpora

We briefly recall general categories and taxonomies of digital forensic corpora within which we embed our work.

Categorization of corpora

Due to their varying nature, corpora in digital forensics fall into different categories, such as disk images, memory images, network packets and files. Corpora of disk images are built with a focus on file system forensic analysis and carving. Memory images are targeting forensic analysis of data structures from operating systems in main memory (RAM). Collections of network packets make the structures and characteristics of communication protocols available, whereas corpora containing files focus on the structures of different application file formats. In this paper, we take a closer look to corpora containing files. Obviously, a corpus specific to SQLite represents a corpus containing database files that conform to the SQLite file format.

Garfinkel describes seven different criteria, that are important when creating forensic corpora (Garfinkel, 2007). They shall help to measure the benefit and usefulness of a corpus. Accordingly, a corpus shall feature the following characteristics.

1. *Representative*: of data encountered during the course of criminal investigations, civil litigation, and intelligence operations.
2. *Complex*: with intertwined information from many sources. Data objects should have a wide range of sizes. The data should be in many human languages.
3. *Heterogeneous*: generated using a range of computer systems and usage patterns.
4. *Annotated*: so that new algorithms can be readily validated against existing ones.
5. *Available*: to researchers operating in an unclassified environment.
6. *Distributed*: in open file formats. Tools should be supplied with the corpora to allow for easy manipulation of the data.
7. *Maintained*: computer systems are constantly changing and evolving. A corpora needs to be routinely augmented with new information or else its interest soon becomes largely historical.

Taxonomy on corpora sensitivity

Additionally, Garfinkel et al. (2009) defined a taxonomy regarding the sensitivity of corpora as follows.

- *Test Data*: Such data sets are specifically created for the purpose of testing and thus do not contain sensitive data. They can be distributed over the Internet without restrictions.
- *Sampled Data*: Data sets that were extracted out of a larger data source, e.g., the Internet. However it might be difficult to ensure that none of the sampled files has legal restrictions regarding redistribution.
- *Realistic Data*: The contents of these data sets can be encountered in real life situations. They can, for example, be collected after installation of software or after having simulated some activities. Distribution is possible from the perspective of privacy, but might be hindered by copyright restrictions.