Computer Standards & Interfaces 32 (2010) 246–255

Contents lists available at ScienceDirect

Computer Standards & Interfaces

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

A proposal for automating investigations in live forensics

Seokhee Lee [a], Antonio Savoldi [b], Kyoung Soo Lim [a], Jong Hyuk Park [c], Sangjin Lee [a,*]

[a] Center for Information Security & Technologies - CIST, Korea University, Anam-dong, Sungbuk-Gu, Seoul, Korea
[b] Department of Electronics for Automation, DEA, University of Brescia, Via Branze, 38, I25123, Brescia, Italy
[c] Department of Computer Science and Engineering, Seoul National University of Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul, 139-742, Korea

ARTICLE INFO

Available online 14 October 2009

Keywords:
Digital evidence collection
Live forensics
Automated digital investigation process
XML technology

ABSTRACT

In this paper we present an XML-based framework, called XLIVE, which provides an efficient way to collect data in live forensic cases, according to well-known crime categories. XLIVE is a forensic automated framework that can be used in live forensic investigations for gathering live data on a Windows-based system. In addition, we have also implemented a proof-of-concept, called LRDS (Live Resource Detection System). This approach of examination will be used extensively to deal with terabyte/petabyte digital systems, where other approaches, such as a post-mortem analysis, cannot be adopted.

Contents

1. Introduction

In this paper we present a new XML-based data collection framework for live forensics purposes which can be used in digital crime investigation [1] for dealing with Windows-based computer systems. It is quite renowned that a typical data collection process is divided into two steps: live data collection and forensic duplication [2]. Live data refers to the digital data which temporarily reside on the volatile memory of the computer being analyzed. Such analysis should provide critical and crucial evidence about the system under examination, such as a “snapshot” at the time of the initial incident response [2]. After the live data collection has been completed, the investigator might perform, according to the well-known paradigms

* Corresponding author.
E-mail addresses: gosky7@korea.ac.kr (S. Lee), antonio.savoldi@ing.unibs.it (A. Savoldi), lukelim@korea.ac.kr (K.S. Lim), parkjonghyuk1@hotmail.com (J.H. Park), sangjin@korea.ac.kr (S. Lee).

0920-5489/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.csi.2009.09.001

of forensic analysis [3], a sound duplication of the hard drive, whether possible, in order to obtain the so-called primary image, which represents a bit-by-bit copy of the whole hard disk of the target system. Furthermore, the investigator will analyze the whole set of data in a forensic laboratory to figure out and possibly solve the digital investigation.

By observing the present trend of state-of-the-art digital crime investigations, we can see that the currently used investigative paradigm is laborious, time-consuming, involves a set of complicated tasks and requires strong skills on the side of the investigator. Therefore, according to the technological developments of storage media, it is clear that the capacity of storage devices, such as hard disks, is increasing more and more, often running into terabytes. Thus, the use of the classical post-mortem analysis approach is becoming problematic especially for large-scale investigations involving a network of computers. In addition, the amount of time available for processing this data is often limited [4].

It is a well-known fact that the digital forensic discipline relies on automation of investigation [5,6], preprocessing for effective searching [4], data collection using profiling methods [7], and data mining techniques [5]. Based on these facts, we could recognize a great demand for automated systems which can support digital crime investigations from an efficiency point of view. Indeed, for this purpose, we have created the XLIVE framework, which meets these features, and is based on three fundamental blocks.

The first one concerns how to gather live data such as specific files and registry information, with regard to the Windows-based computer category. This approach can help investigators in circumstances when it is impossible to perform a forensic duplication on the target system, as illustrated below:

Rapid incident response: the forensic practitioner has no time to create a bit-by-bit digital image, the so-called primary copy;
Massive storage system: the size of the storage system is too large (for instance, a data center based on a petabyte storage system);
Mission critical systems: such computer systems cannot be switched off for the criticality of the service being done. Thus, the only feasible approach from a forensic point of view is to use a proper live analysis methodology [8].

The second block of the framework concerns how to collect digital objects automatically, on the basis of certain crime categories, which need to be created according to specific criteria. Thus, according to specific profiles, only selected digital objects will be gathered, which are those that are possibly related to the crime we are investigating. This approach can speed up the investigation and can be adopted even by a non-expert forensic practitioner.

The third block regards how to organize documents using the well-known XML format according to specific templates. Moreover, the whole framework is based on the XML technology, which has already been greatly used for other forensic systems and is a *de facto* standard for describing evidence [4]. This fact implies several advantages such as having a tree-based structure format, an intermediate format of data processing, and a simple way to organize data workflows [9].

These techniques are combined in a proof-of-concept named LRDS (Live Resource Detection System), and can be used to collect digital evidence from Windows-based live systems.

2. XML technology and an excerpt of live data

The eXtensible Markup Language (XML) is a consolidated standard language which can be used to describe raw data, such as those that are collected during a digital investigation. Fig. 1 illustrates an excerpt of wrapped data which is the result of the `netstat -an` command. All relevant data, such as the protocol, the IP address and the state of the connection are well indicated by the XML tag. Every XML code snippet, which is delimited by `<connection>` tags, refers to a precise portion of the mentioned command output. As a consequence, the results can be read quite easily because of the tree-based structure format and the tags which define the meaning of the raw content. Afterwards, XSL (Extensible Stylesheet Language) allows to transform an XML document into another one, according to the different profiles and crime categories, which are created by means of profiling science [10]. This facilitates forensic experts, analysts, attorneys, and judges to understand the collected information [11].

2.1. Related work

There are several ongoing projects that aim at automating digital investigations involving large evidence sets. Unfortunately, most of them are focused on post-mortem analysis [1,12–14]. One of the projects which has already been implemented is the Phisherman Project [15], which is based on the well-known IODEF format (Incident Object Description and Exchange Format). IODEF has been defined by computer security incident response teams (CSIRTs) to exchange operational and statistical incident information among themselves, their constituencies, and their collaborators. It can also provide the basis for the development of an interoperable set of tools and procedures for incident reporting [16].

The IODEF format aims to:

- increase automation in the processing phase of an incident, since the workload of the security analyst will be reduced;
- decrease efforts in normalizing similar data from different sources, even though they are well structured;
- define a common format which can be used to build interoperable tools to be used for incident handling and the subsequent analysis phase. This can be extremely useful when data comes from multiple constituencies.

Moreover, this format is being reviewed by IETF to define an international standard. Fig. 2 shows an excerpt of the above-mentioned format. It refers to a phishing website whose IP address and fraud type are shown.

2.2. XIRAF

XIRAF (XML-based indexing and querying for digital forensics) describes an XML-based framework to manage and query forensic traces extracted from digital evidence. XIRAF systematically applies forensic analysis tools to evidence files (e.g. hard disk images). Each tool produces structured XML annotations that can refer to regions (byte ranges) within an evidence file. XIRAF stores such annotations in an XML database, which allows us to query the annotations using a single, powerful query language named XQuery [4]. The XIRAF framework consists of three main components. The tool repository houses a collection of feature extraction tools. The feature extraction manager orchestrates the invocation of these tools, merges their XML outputs, and stores the result in the storage subsystem. The storage subsystem consists of binary large objects that hold raw evidence data and an XML database that holds all extracted features. Certainly, this framework can enhance the reconstruction of digital criminal events which can usually involve taking care of plenty of digital objects.

2.3. FACE

FACE [17] is a framework for automatic evidence discovery and correlation from a variety of forensic targets. The main purpose of this framework is the analysis and correlation of a disk image, memory image, network capture, and configuration log files. Thus, it can be