On Reducing Bottlenecks in Digital Forensics

by Martin Schmiedecker and Sebastian Neuner (SBA Research)

Digital rensic investigators currently face numerous challenges, some of which include: the increased digitalisation of our lives, vast case sizes owing to ever increasing storage capacity and the large number of personal devices in use. The goal of our SpeedFor project is to develop new methodologies to reduce the manual work required for digital investigators. Among other things we harvest information from file sharing networks to identify files by extending the forensic process. In an initial proof-of-concept we obtained information from the BitTorrent network to identify up to 2,500 terabytes of data.

Digital forensics has received increasing attention in recent years, as more and more crimes are conducted exclusively by, or with the involvement of, computers. Tools and methods of digital forensics are used by private investigators as well as law enforcement analysts all over the world. One of the challenges in digital forensics is the vast amount of data that needs to be analysed. Commodity hard drives with eight terabytes and more storage capacity are standard nowadays, and are readily obtained. Current forensic processes, however, do not scale well to multi-terabyte workloads. Our project SpeedFor aims to fundamentally increase the performance of current state-of-the-art forensic methods and decrease the manual work necessary for a forensic analyst by developing new methods to increase the use of parallelised data processing within the specific environment of digital forensics, and identifying the best methods to exclude a possibly vast number of files and file system artefacts that are not specific to a case.

We aim to achieve these objectives by leveraging the information specific to large file-sharing networks, in particular we use the BitTorrent protocol as source of information. Prior to sharing a set of files in the network, the data is split in to ‘chunks’ to facilitate parallel data transfers. These chunks are then hashed, and stored in the meta-information of the Torrent swarm [1]. By crawling popular torrent swarms, as well as the few remaining BitTorrent websites, the computational power of the initial seeders can be used for good i.e., a methodology to identify files and file fragments based on data from publicly available file-sharing networks.

Our prototype, dubbed peekaTorrent, will be published at the upcoming DFRWS conference. peekaTorrent is based on the open-source forensic tools bulk extractor [2] and hashdb [3], and can be readily integrated into forensic processes. It improves the current state of the art on sub-file hashing twofold: compared to previous approaches we hash sub-file parts larger than pure sector-based hashes. Figure 1 shows a graphical representation. Previously the hard drive sectors were used as input for hash functions like SHA-1, shown as the second layer on the bottom of Figure 1 for hard drives which use 512 bytes respectively 4K sectors. We propose to extract the already available data using larger hash windows which overlap with the BitTorrent specification. Our method is less prone to false-positives for files that share common data seg-

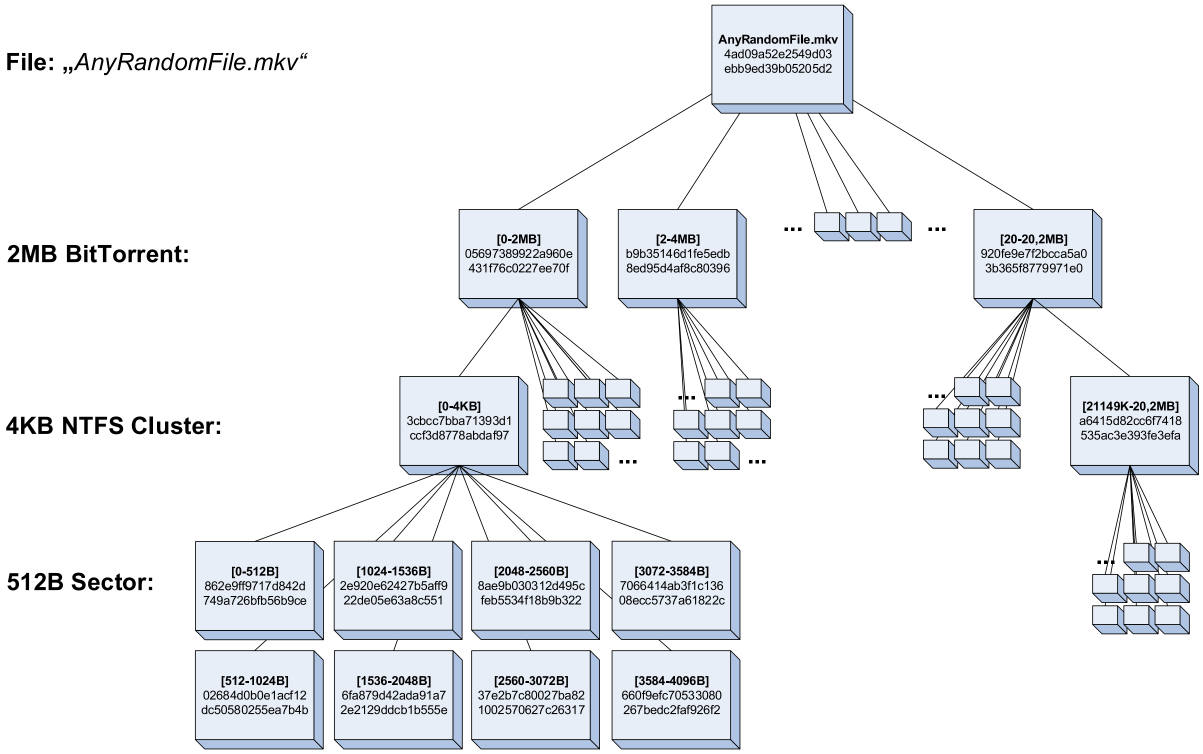

Figure 1: Sub-file hashing as used in peekaTorrent.

ments due to the larger area to be hashed.

Secondly, we solve the problem of an a-priori sub-file hash database being required by creating one that can be shared openly. We collected more than 2.5 million torrent files and built their corresponding hashdb databases, which are freely available on the project website. Note that no participation in file-sharing activity is needed as the torrent metadata or 'metainfo' already contains all the necessary information including the sub-file hash values. We publish all data sets, tools and our paper openly – you can find the source codes and the hashdb files released under open-source licence on our website https://www.peekatorrent.org. Overall, we expect these methods to help investigators around the globe as this information can be used for file and file fragment identification as well as for effective file whitelisting.

Link:
[L1] https://www.peekatorrent.org

References:

[1] Bram Cohen. The BitTorrent Protocol Specification, BEP-3, online, http://www.bittorrent.org/beps/bep_0003.html

[2] Simson L. Garfinkel: "Digital media triage with bulk data analysis and bulk_extractor", Computers & Security 32: 56-72, 2013

[3] Simson L. Garfinkel, Michael McCarrin: "Hash-based carving: Searching media for complete files and file fragments with sector hashing and hashdb", Digital Investigation 14: S95-S105, 2015.

Please contact:

Martin Schmiedecker
SBA Research, Austria
mschmiedecker@sba-research.org

Multi-View Security and Surveillance at MTA SZTAKI

by László Havasi and Tamás Szirányi (MTA SZTAKI)

The Distributed Events Analysis Research Laboratory (DEVA) has more than 10 years of research experience in security and surveillance, including multi-view systems of optical, thermal, infra-red and time-of-flight cameras, as well as LIDAR sensors. The laboratory's research and development work has been addressing critical issues of surveillance systems regarding the protection of critical infrastructures against incursions and terrorist attacks.

The DEVA laboratory has been involved in several security related projects funded by the European Commission and the European Defence Agency, contributing significant improvements regarding multi-view computer vision and target tracking efforts. During a recently finished project (PROACTIVE, EU FP-7 [L1, L2]) that included several European partners in defence and security, a holistic IoT framework was developed enabling enhanced situational awareness in urban environments in order to pre-empt and effectively respond to terrorist attacks. The framework integrates many novel technologies enabling information collection, filtering, analysis and fusion from multiple, geographically dispersed devices. At the same time, the framework integrates advanced reasoning techniques in order to intelligently process and derive high level terrorist oriented semantics from a multitude of sensor streams.

The DEVA Laboratory is responsible for processing and understanding multimodal visual information from cameras and 3D sensors, sampled in different time instants, and situated in different locations. Special emphasis is on the fusion of different sources, such as satellite or airborne image data for remote sensing, potentially amended with terrestrial and UAV based imaging.
In our surveillance projects, an important issue is the tracking of objects/targets, and the detection and recognition of events by using multi-view camera networks, including infra-red sensors. In these applications calibration is always a problem, since security scenarios usually require quick installation and continuous troubleshooting. Another challenge is the co-registration of optical cameras and infra sensors for 3D tracking, since features of different modalities are usually hard to associate and compare.

In a recently finished project (PROACTIVE, EU FP-7) that included several European partners in defence and security, our main task was the visual tracking and analysis of human and vehicle behaviour [1] and crowd events.

The project addressed some specific emergency situations involving man-made or natural disasters and terrorism, i.e., frequent threats within our society. Avoiding an incident and mitigating its potential consequences requires the development and deployment of new solutions that exploit the recent advances in terms of technological platforms and problem solving strategies.

PROACTIVE produced an end-user driven solution. PROACTIVE prototypes include the following parts:

- *Terrorist Reasoning Kernel:* the reasoning layer provides the needed intelligence in order to infer additional information regarding the incoming suspicious event stream. This layer aids law enforcement officers by reasoning about threat levels of each incoming event and potentially inferring its association with a possible terrorist attack.
- *Context Awareness Kernel:* these processing modules provide semantic description about the environment and the static and moving (e.g., foreground) objects and sufficient information about the suspicious events and actions.
- *C2 platform:* the command and control platform is a multi-touch and multi-user web-application that provides a graphical user interface and enables the user to view maps (2D/3D), devices/sensors and alerts from the system.