B21 Bulk Data Analysis With Optimistic Decompression and Sector Hashing

Simson L. Garfinkel, PhD, 1186 N Utah St, Arlington, VA 22201; Kristina Foster, MS, Naval Postgraduate School, 900 N Glebe, Arlington, VA 22203; Joel D. Young, PhD, Naval Postgraduate School, 1 University Cir, Monterey, CA 93943; and Kevin Fairbanks, PhD, Naval Postgraduate School, 900 N Glebe, Arlington, VA 22203*

After attending this presentation, attendees will understand the difference between file-based and bulk-data approaches to digital forensics, and how bulk data approaches can be significantly empowered through the use of optimistic decompression and sector hashing.

This presentation will impact the forensic science community by providing results from controlled experiments in an area with little previous research, adding to work being carried out in digital forensics by broadening the understanding of the presence of compressed data, especially fragmented compressed data, in unallocated areas of file systems, and the availability of new techniques to decompress such data without reference to file system metadata or file type. This presentation will also build upon previous research in sector hashing (also known as piecewise hashing and with hash-based carving and show how sector hashing can be empowered through the use of a high-performance custom-built database.[1-5]

Bulk data analysis is a new digital forensics technique that eschews file extraction, and instead focuses on the processing of bulk data read directly from the target media. Unlike file-based approaches, bulk data analysis is particularly well suited to triage, as it can be parallelized and applied to random sampling.

In the first experiment, a corpus of roughly 2,000 hard drives purchased on the secondary market was analyzed for forensically important information that could only be recovered through the use of optimistic decompression of sectors that were not contained within allocated or recoverable deleted files.[6] Optimistic decompression means that all decompression algorithms are applied to all sectors with the hope that some compressed data may be identified and decompressed. This experiment employed the use of the bulk-extractor, fiwalk and identify-filenames.py tools.[7,8] This study found a significant number of email addresses, URLs, account numbers, and other kinds of information that could only be recovered through the use of optimistic decompression. This presentation show how additional processing can be used to determine which recovered features are attributable to user-generated content, and which are residual data from software distributions.

In the second experiment, three corpora (GOVDOCS1, OpenMalware 2012, and NSRL) were hashed on sector boundaries and evaluated for the presence of distinct sectors—that is, sectors which are not present in any of the sectors in the corpus.[6,9-11] Many such distinct sectors are present and how they can be used to identify the presence of either intact files or to attribute residual data to specific files of interest will be shown.

In conclusion, these two studies provide evidence that bulk data processing can be productively used by digital forensics examiners for both triage and for the extraction of case-relevant details. Examiners can use the tools presented in this presentation today. This presentation will also provide sufficient information so these techniques can be embedded into other tools.

References:

[1] Nicholas Harbour. dcfldd, 2006. http://dcfldd.sf.net.

[2] Jesse Kornblum. md5deep and hashdeep—latest version 4.1, June 26 2011. http://md5deep.sourceforge.net/. Last accessed Feb. 18, 2012.

[3] Simson L. Garfinkel. announcing frag_find: finding file fragments in disk images using sector hashing, March 2009. http://tech.groups.yahoo.com/group/linux_forensics/message/ 3063.

[4] Yoginder Singh Dandass, Nathan Joseph Necaise, and Sherry Reede Thomas. An empirical analysis of disk sector hashes for data carving. Journal of Digital Forensic Practice, 2:95–104, 2008.

[5] Sylvain Collange, Marc Daumas, Yoginder S. Dandass, and David Defour. Using graphics processors for parallelizing hash-based data carving. In Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009. http: //hal.archives- ouvertes.fr/docs/00/35/09/62/PDF/ColDanDauDef09.pdf. Last accessed Dec. 3, 2011.

[6] Simson L. Garfinkel, Paul Farrell, Vassil Roussev, and George Dinolt. Bringing science to digital forensics with standardized forensic corpora. In Proceedings of the 9th Annual Digital Forensic Research Workshop (DFRWS). Elsevier, Quebec, CA, August 2009.

[7] Simson Garfinkel. Stream-based digital media forensics with bulk_extractor. 2012. In Submission.

[8] Simson Garfinkel. Digital Forensics XML. Digital Investigation, 8:161–174, February 2012. Accepted for publication.

[9] Danny Quist. State of offensive computing, July 2012. http://www.offensivecomputing.net/?q=node/1868.

[10] National Institute of Standards and Technology. National software reference library, March 2012. http://www.nsrl.nist. gov/.

[11] Simson Garfinkel, Alex Nelson, Douglas White, and Vassil Roussev. Using purpose-built functions and block hashes to enable small block and sub-file forensics. In Proc. of the Tenth Annual DFRWS Conference. Elsevier, Portland, OR, 2010.

Digital Forensics, Optimistic Decompression, Sector Hashing

** Presenting Author*