MemCorp: An Open Data Corpus for Memory Analysis

Timothy Vidas
Carnegie Mellon University
tvidas@cmu.edu

Abstract

Memory analysis, as with most areas of digital forensics, has suffered from a lack of open and available datasets. Such sets could be used to facilitate collaboration among researchers and practitioners, validate previous results or assess the capabilities of a tool. A memory analysis corpus may also be used by educators in a multitude of academic settings. This paper explores the needs in research, practical and educational settings and introduces such a corpus.

1. Introduction

Science can broadly be thought of as systematically acquired and verifiable knowledge. Similarly forensic science is the application of different fields of science to a legal system. In the budding field of digital forensic science, there is currently very little capability to verify results. Many results are published in the form of blog posts, conferences that have no proceedings or as ancillary footnotes in books. These results along with those published in academic conferences and journals often include not only new software and methods, but also the creation of new datasets upon which to act. In order for another party to verify conclusions the bar is set high. The other party must procure valid software and data as well as following the proper process.

In many cases there is no reason that the data used in a research project need be unique. The data is simply created because there is no ready access to acceptable data.

Without data that is pertinent to the experimentation that preceded a conclusion, it may not just be difficult but it may be impossible to reproduce results.

MemCorp is made available to researchers, educators and practitioners alike in order to provide a common dataset for use with education and research in the area of memory analysis.

2. Related Work

There has long been a lack of available datasets in the digital forensic community. Focused examples can found in the Digital Forensics Tool Testing [2] and NIST's CFREDS projects [14]. Recently the Real Data Corpus (RDC)[10] has made large quantities of *real* data available subject to certain restrictions on some portions of the RDC (see [10] or http://digitalcorpora.org/corpora/disk-images/rdc-faq for more information). Both CFREDS and RDC include a limited, small quantity of memory images.

In an ad-hoc manner, various digital forensics competitions have made data files publically available. In particular, annual challenges found at DFRWS[7], the DoD Cyber Crimes Conference [5], Defcon's Capture the Flag qualification round[6] and NYPolytechnics Cyber Security Awareness Week[16] have provided some public data files that, possibly contrary to their original purpose, can and are used to facilitate research and education. In fact the first three have all made some samples specifically related to memory analysis publically available.

The images available through MemCorp comprise the most comprehensive set (by both quantity and diversity) of memory images publically available via the above mentioned resources, and are available to any party without restriction.

Outside of digital forensics, there are numerous examples of corpora used to verify results, limit the effort required to start a research project and in general to simply facilitate collaboration. Among these corpora are instances in network traffic [15], email [12], news [13], search[1] and the list goes on.

3. Benefits of a Public Corpus

Forensics involves the application of a science to a legal system, as such forensics research most often falls into the realm of applied research (as opposed to basic research). However, in many cases digital forensics researchers may be disconnected from

practitioners who habitually have immediate needs for specific results. A public corpus can not only serve each group independently, but can also act as a catalyst for encouraging collaboration between the two.

3.1. Practitioner Needs

Practitioners need ways to verify the advertised capabilities of the tools they use. Memory collection and analysis is becoming more mainstream, but tools are still laden with specific use cases, numerous exceptions to the rule and limited applicability. For example, the open source Volatility Framework [22] only supports the analysis of Windows XP (service pack 2 or 3). Since some tools have severe limitations or specific targets, it is import for public datasets to be diverse enough to support a wide range of cases.

Similar to the verification of capabilities, the user of a tool should be able to periodically confirm that the tool continues to perform as desired. This confirmation should be easily performed over time and across tool versions.

When selecting the appropriate tool for a memory analysis task, the practitioner is currently largely left to chose based on vendor reputation or to create datasets and perform a large amount of testing. Tool comparison should not be limited by the availability of applicable datasets.

3.2. Researcher Needs

Not only do researchers often have the ability to build on related work, but the duplication of existing work is frequently viewed negatively as an ignorance of the existence of previous work, or worse, as a blind omission. In either case the new work is regarded as un-needed and at best is cited as confirmation of previous results. In memory analysis, results can be difficult to recreate and thus improve upon simply because datasets the original researcher used are unavailable or even lost. The lack of data causes new researchers to re-create data and then to re-create results simply to obtain a starting condition for their own new research. The re-creation of data results in a large expenditure of effort for little gain and the use of a public corpus by both researchers greatly reduces the amount of replicated effort.

Even researchers starting an entirely new course of investigation benefit by having a baseline with which to start, saving them the effort of dataset creation as mentioned above. Researchers with similar objectives can use a shared dataset for collaboration opportunities. Numerous research papers relating to memory forensics state or imply the creation of datasets [8,9,17-21].

Researchers with goals of solving "real problems" have a need to interface with practitioners who interact and participate with "real problems" regularly. As with other computer security related fields, the sensitivity of real data varies based on intent to prosecute legally, non-disclosure, government classification, and other similar restrictions. In this case a well crafted, openly available corpus may be able to offer surrogate data that the practitioner and researcher can use as common ground.

3.3. Educator Needs

Like the practitioner and researcher, a digital forensics educator suffers from similar effort expending disadvantages related to dataset creation. Furthermore it is difficult to craft consistent samples, as the classes are commonly structured toward teaching students to find irregularities and anomalies. Even the aforementioned digital forensics challenges contain unintended information and other "mistakes."

A publically available dataset can greatly reduce the current requirement of initial effort on the part of an educator. Similarly the educator can take advantage of different parts of a corpus over time reducing the amount of effort that would otherwise be spent on maintaining their own dataset. Akin to the researchers, educators have begun using publicly available portions of forensics challenges in the classroom [3].

4. MemCorp Composition

The memory image corpus outlined here takes several types of memory images into account. Foremost the corpus serves to provide a common baseline removing the need for others to duplicate the effort of image creation, but the corpus strives to address all the needs outlined above. A substantial part of the corpus was collected using virtual machines, but it is recognized that a certain amount of realism and diversity would be lost if the corpus was created solely from virtual machines, and as such images were also acquired directly from physical machines. Some research was done to determine the best tool for performing physical collection, since that work has already been performed, these images

are also made available as part of the corpus (see section 5 for more details).

The corpus currently consists of 87[1] Microsoft Windows based images ranging from Windows 95 to Windows 7 (See Figure 1). The distribution of images by base operating system is outlined in Figure 2.

Figure 1: Corpus Distribution by Collection Type	
Virtual Machine	53
Physical Machine	23
Case Study	11

Figure2: Corpus Distribution by Operating System	
Windows 9x	3
Windows NT	2
Windows 2000	12
Windows 2003	15
Windows XP	19
Windows Vista	18
Windows 2008	6
Windows 7	12

4.1. Virtual Machine Images

Virtual machines have the need to store "virtual physical memory" for the virtual machine. That is memory that the virtual machine "sees" as physical RAM. Predominately this memory is a block of physical memory on the host machine, however when a virtual machine is paused or suspended with the intention of resuming the virtual machine at a later point in time, the state of this memory (and disk, etc) must somehow be saved. Some virtualization solutions implement the saving of virtual RAM to a single file representing physical address space. This allows for the "collection" (eg. copy of a file) to happen instantaneously with respect to the virtual machine. Images collected in this way will have no side effects that may be introduced by a memory collection tool, and will exhibit characteristics of a physical machine to the extent that the virtualization environment can replicated the characteristics.

Images in the corpus were collected from the host file system after the operating system was installed onto the virtual machine. Immediately after operating system install, the virtual machine was rebooted and then paused yielding a file with the contents of physical memory.

[1] There are also 31 additional images available as a result of the case study detailed in Section 5. These 31 images from the same base virtual machine are not included in Figures 1 or 2.

4.2. Physical Machine Images

Acquiring images from non-virtualized systems adds considerable diversity to the corpus. Since these images are collected from physical machines using practitioner techniques, these images exhibit artifacts created by the collection tool as well as characteristics introduced by the hardware configuration.

In an effort to keep the corpus openly available, and to facilitate the use of this portion of the corpus as a baseline for collected image analysis, these images were acquired immediately after the initial unpacking of a system from the vendor. This collection strategy also eliminates the possibility of introducing any PII (personally identifiable information) into this part of the corpus.

4.3. Case Study Images

The base cases provided by the virtual machine and physical machine images may be too uniform for some applications. Case study images offer images collected from fabricated situations (such as a machine with active network connections or with malware present).

Since every possible future use case cannot be envisioned, the corpus includes a range of memory images that exhibit qualities that are known to work with some tools, and qualities that are thought to possible. In areas where the corpus found to be deficient, the corpus can always be extended to include new cases.

The main intent of this portion of the corpus is to provide images are that are not base cases and can be used to demonstrate and research capabilities that are incompatible with base cases. Educators may find images in this category useful for classroom use since these images allow for more avenues of analysis, but are still created in a controlled manner.

4.4. "Real World" Images

Given that memory images must generally be acquired from a running machine, it is difficult to gather images without cooperation from other parties. Persons, corporations, and institutions have to deal with issues of privacy and secrecy when considering the possibility of making a data store such as a memory image publically available. Furthermore, introducing such memory images allows for the inclusion of PII which may affect the dissemination and usability of this portion of the corpus.

Even with the above concerns, *real* data frequently manifests conditions that are unlikely to

ever be created in a controlled manner such as the case study images. MemCorp doesn't currently contain any real world images, and the reader is referred to the "boomer" image in the CFREDS dataset or the M44 set from the RDC.

5. Memory Collection Tool Study

In order to determine the appropriate methodology for collecting images from physical machines a small study was performed. The results of this study are presented here, largely so that others may build from this work and secondarily to support the selection of win32dd (and win64dd) as the tool of choice for collecting memory from physical machines for inclusion in MemCorp.

5.1. Tool selection process

VMware Workstaion was used to create a virtual machine (with no network card). A baseline snapshot was created in VMware. All the tools to be tested were executed or installed from the same USB media. The USB media was connected to the virtual machine. If required, the tool was installed. A memory image was collected from the virtual machine (to the USB media). Once the imaging had completed another snapshot was created. At this point in time, for each tool there are three data points for analysis: the baseline (via snapshot), the image collected using the tool, and the state of memory after the tool concluded (via snapshot).

The three images were compared at the page level to estimate the impact that simply running the tool had on memory contents and to observe the accuracy of the collected image.

Additionally the virtual machine was allowed, starting at the baseline snapshot, to run independent of any collection tools. Snapshots were created at various intervals to determine a reference of change that would occur without the introduction of any collection tools. These reference images allow for similar comparisons of "idle operation" over time which can account for some amount of memory state change due to normal background processes.

Note that previous research argues that imaging across a network socket likely has less system impact than imaging to USB. However, in the authors experience, the use of USB is far more likely to occur in the field with practitioners and as such, the collection process was performed via USB.

5.2. Tools tested

The following memory collection tools were considered in this study: KntDD, win32dd (and win64dd), mdd, Nigilant32, Prodiscover IR, memoryze, FastDumpPro, fastdump community, f-response + ftkimager, ftkimager, winen. Since tool comparison is not the focus of this paper, only brief comments are only made on tools that demonstrated operation outside of the norm or otherwise warranted specific attention:

A) KntDD operated at about 1% per minute over 1GB of RAM. This is likely too slow to be used operationally.

B) Mantech's mdd product allocates the entire destination image immediately so file size cannot be used to gauge progress, however there is a progress bar in the window title bar (not in the text portion of the window). mdd is known not to work on some versions of windows, and not work on installations with more than 4 GB of memory. Perhaps more importantly, mdd had the largest impact to memory, by far.

C) In many cases Nigilant32 fails to run, even as administrator and is unable to open the PhysicalMemory object.

D) Prodiscover IR is estimated to cost approximately $13000 USD and as such was not tested.

E) Memoryze requires installation, which increases the impact this tool has on the host machine. This limits its usefulness in many first responder and digital forensics collection scenarios.

F) FastDump community is remarkably slow, so slow it is unlikely the tool would ever be used in practice.

5.3. Memory Collection Study Results

win32dd and win64dd[2] were selected primarily due to observed correctness and minimal impact to memory and secondarily due to speed consideration. Figure 3 shows the difference between the before and after snapshot images, which can be thought of as "the impact that running the tool had on memory." It is easy to see that win32dd is slightly better than other competitors (regardless of which win32dd options were employed) and that mdd is by far the worst.

[2] Note: after this study was performed, Moonsols released new versions of win32dd and win64dd. With regard to this study, cursory tests showed similar results to the original non-Moonsols versions.

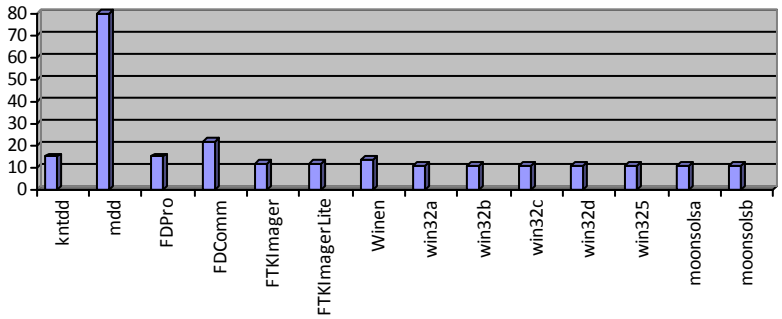

Figure 3: Percent memory changed as a result of running the tool.

6. Conclusions and Future work

The memory analysis corpus presented here is currently available at no cost to any person that wishes to download the corpus. MemCorp is structured to meet current and some future needs of digital forensics practitioners, researchers, and educators.

NIST has agreed to host MemCorp datasets along with sample analysis data. As the data included in MemCorp is not *real* it is unlikely that MemCorp will ever be included in the Real Data Corpus. Regardless, MemCorp is currently available through the author's website [4].

A logical next step is to include memory images for other operating systems, such as MacOS or Linux. The largest issue faced with such an addition is considerable increase in volume. In unix based operating systems different distributions, or even builds employ varying memory structures making it difficult to capture every possible case. The addition of these types of images will be reconsidered as the evolution of non-Windows memory forensics matures.

Since many virtual machines use hardware assisted virtualization (as opposed to para-virutalization) the memory layouts may actually be different when the same virtual machine is deployed on different hardware. Some subset of the virtualized images, may be duplicated to incorporate hardware changes.

7. References

[1] Brants, T. and Franz, A. "Web 1T 5-gram corpus version 1.1", Linguistic Data Consortium, Philadelphia, 2006.

[2] Carrier, B. Digital forensics tool testing images, http://dftt.sourceforge.net, 2007.

[3] Collins, D and McGuire, T. "Using the DC3 forensic challenge as a basis for a special topics digital forensics upper level undergraduate course", Journal of Computing Sciences in Colleges 23. Number 6, 2008, pp8-14..

[4] CMU MemCorp webpage. http://www.ece.cmu.edu/~tvidas/

[5] DC3 Digital Forensics Challenge. http://www.dc3.mil/challenge.

[6] Defcon Capture the Flag Archive. https://www.defcon.org/html/links/dc-ctf.html.

[7] Digital Forensics Research Workshop forensic challenges. http://www.dfwrs.org.

[8] Dolan-Gavitt, B. "A process-eye view of physical memory", Digital Investigation 4, 2007, pp. 62-64.

[9] Dolan-Gavitt, B. "Forensic analysis of the Windows registry in memory", Digital Investigation 5, 2008, pp. S26-32.

[10] Garfinkel, S. Farrel, P. Roussev, V and Dinolt, G. "Bringing science to digital forensics with standardized forensic corpora", Digital Investigation 6, 2009, pp. S2–S11.

[11] Halderman J., Schoen S., Heninger N. et al. "Lest We Remember: Cold Boot Attacks on Encryption Keys" USENIX Security '08 proceedings. 2008. pp 45-60.

[12] Klimt, B. and Yang, Y. "Introducing the Enron corpus", First Conference on Email and Anti-Spam (CEAS), 2004.

[13] Lewis, D. D., Yang, Y, Rose, T. and Li, F. "RCV1: A New Benchmark Collection for Text Categorization Research", Journal of Machine Learning Research 5, 2004, pp. 361-397.

[14] Lyle, J. The CFREDS project. http://www.cfreds.nist.gov. 2008.

[15] Mawi working group traffic archive, http://tracer.csl.sony.co.jp/mawi, 2009.

[16] NY Poly Cyber Security Awareness Week website. http://www.poly.edu/csaw

[17] Petroni, N et al. "FATKit: A framework for the extraction and analysis of digital forensic data from volatile system memory", Digital Investigation 3, 2006, pp197-210.

[18] Schuster, A. "Searching for processes and threads in Microsoft Windows memory dumps", Digital Investigation 3, 2006, pp. 10-16.

[19] Schuster, A. "Pool allocations as an information source in Windows memory forensics", International conference on IT-incident management and IT-forensics, 2006, pp. 104-115.

[20] Van Baar, RB and Alink, W. and Van Ballegooij. "Forensic memory analysis: Files mapped in memory", Digital Investigation 5, 2008, pp. S52-57.

[21] Vidas, T. "The acquisition and analysis of random access memory". Journal of Digital Forensic Practice 1. Dec. 2006, pp 315–323

[22] Volatility Framework. Volatile Systems. https://www.volatilesystems.com/default/volatility