Column:
Factors Affecting Data Decay

Kevin Fairbanks
Johns Hopkins University Applied Physics Laboratory

Simson Garfinkel
Naval Postgraduate School

In nuclear physics, the phrase *decay rate* is used to denote the rate that atoms and other particles spontaneously decompose. Uranium-235 famously decays into a variety of daughter isotopes including Thorium and Neptunium, which themselves decay to others. Decay rates are widely observed and wildly different depending on many factors, both internal and external. U-235 has a half-life of 703,800,000 years, for example, while free neutrons have a half-life of 611 seconds and neutrons in an atomic nucleus are stable.

We posit that data in computer systems also experiences some kind of statistical decay process and thus also has a discernible decay rate. Like atomic decay, data decay fluctuates wildly. But unlike atomic decay, data decay rates are the result of so many different interplaying processes that we currently do not understand them well enough to come up with quantifiable numbers. Nevertheless, we believe that it is useful to discuss some of the factors that impact the data decay rate, for these factors frequently determine whether useful data about a subject can be recovered by forensic investigation.

Computer systems have grown so reliable in recent decades that most of today's users cannot remember a time when printouts were the only form of persistent storage available to most users. Today, information stored on a hard drive can almost always be retrieved months or even years later unless it is intentionally overwritten. This was not the case in the 1970s, when mass storage systems were subject to frequent failures.

The retrievability of data written to mass storage is a function of many factors including the reliability of the underlying storage system, the correctness of the computer's file system implementation, the file system data structures, whether the data remain allocated or are "deleted," and usage patterns. As hardware has become more reliable, non-hardware failures have come to dominate. Indeed, different implementations of the same file system may have widely different data decay rates for both allocated and deleted data. Journaling file systems are more likely to retain allocated data than file systems that do not employ journals. But different implementations also impact the decay rate of deleted data: Microsoft's implementation of the FAT file system deletes a file by changing the first byte of the file name to 0xE5 and putting the file's clusters on the free list, while other implementations overwrite the entire directory entry so that the name cannot be recovered. Thus Microsoft's decay rate for deleted FAT data is lower than those of other implementations.

Because differing implementations result in differing data decay rates, it is difficult *a priori* to predict the success rate of a particular recovery effort without knowing a vast number of case-specific details. Nevertheless, we argue that the "data decay rate" is a useful concept for the forensic examiner to entertain. Estimating the likelihood of data recovery is an essential part of triage. Given limited forensic resources, an estimate of the decay rate may be an important part of prioritizing an investigation. Understanding the decay rate may also help an examiner determine if data were erased during the course of normal system operations or if there was an intentional effort to destroy evidence.

Allocation Status: This is clearly one of the most important factors impacting the rate of data decay. Properly functioning systems will attempt to preserve allocated data unless it is intentionally overwritten; unallocated or deleted data may be overwritten as a result of normal system operations. Most file systems on rewritable media will reuse sectors associated with a file when the file is erased.

Resue Policy: Once a file is deleted, if there are no other hard-links to the data, the addresses of the data units are marked as available for usage. The reuse policy will significantly impact the data decay rate.

Media Type: This is also an important factor in data decay. For example, RAM is highly volatile during the normal operation of a computer: we would expect deleted data in RAM to have a significantly faster decay rate than deleted data on a magnetic drive. Meanwhile, deleted data on write-once media (e.g., CDRs) will likely persist even after it is no longer visible to the user, and be recoverable by forensic tools.

Data Type: Different kinds of data have different decay rates. File system metadata, such as timestamps, or data manipulated by an application that an end user regularly uses, such as an email database, are likely to change much faster than system libraries and executables.

Media Degradation: These are errors caused by some event, including the passage of time. For example, data on magnetic media would decay faster when in strong electro magnetic fields, while data on optical media would not.

Media Errors: This includes errors introduced during writing, correctable transient errors, and uncorrectable permanent errors. High error rates may result in high data decay rates.

File System Design: Many decisions made by the designers of the file system will impact the data decay rate, including the strategy that the system uses to allocate blocks for new files, whether or not blocks are preallocated, and the frequency of defragmentation. Even the implementation of the defragmentation process complicates matters: Historically defragmentation was performed by a specialized tool while the file system was offline; newer systems can perform defragmentation during normal operation. More frequent defragmentation likely increases the decay rate.

Fragmentation Issues: While fragmentation and preallocation can overwrite deleted data, there are situations where they can also result in deleted data being recovered. If a file is created in a fragmented fashion, defragmented, and then deleted, then there exists a possibility that a portion of its contents remains in the original data blocks. The probability of complete recovery of the file contents will depend on the tool that performed the defragmentation and the location on the disk from which the fragmented blocks were moved.

Preallocation Strategies: The preallocation routines in the ext4 file system may also result in paradoxically lowering the data decay rate, as ext4 does not clear the blocks used for preallocation, so these blocks will essentially be frozen with their previous contents. Again, depending on the size of the deleted content, complete recovery may not be possible: however, if the examiner is seeking to prove that a known file did reside on the target disk at one point in time, then techniques such as sector hash comparison can be undertaken.

File System Size, Utilization, and Age: As a file system ages, large spans of contiguous blocks can become less frequent. In this environment, the possibility of file fragmentation increases depending on the file size. It also stands to reason that larger files will take longer to be completely overwritten, thereby resulting in a higher probability of partial recovery. With some types of files, such as encrypted files, recovering a single block of data can be used as evidence that the entire file was once present. This may be significant in an investigation.

Drive Usage Patterns: If the drive that contains the deleted data also contains system files for the operating system, software updates may increase the data decay rate. Furthermore, users generally use their personal computers for more than one purpose. This results in a mixture of archetype behaviors such as a downloader, a browser, and a gamer all within the same profile on one system. This situation may be exacerbated on mobile devices such as smartphones and tablets, as these machines have relatively small amounts of storage and are in constant use.

Implementation Errors: Experimental or new file systems are likely to have bugs. Thus data stored on them is likely to decay faster than data stored on better understood systems.

Clear definitions and procedures for measuring the data decay rate of particular systems will help improve our understanding of digital forensics. Is it reasonable to define a data recovery rate as the percentage of blocks that can be recovered after a certain amount of time or operations? If 50 out of 100 blocks of a PDF file can be found in an unallocated portion of a hard drive image, can this information be used to determine when the PDF file was deleted -- or if its deletion was the result of an intentional act or an automated system process?

In order to understand and quantify how deleted data decays, one must understand file system and user behavior. Analyzing the implementation of a file system will allow the identification and classification of volatile fields during normal file system operation. We can use simulation and our understanding of block allocation, preallocation, and fragmentation handling to develop probabilistic models for data decay. Furthermore, user behavior can be understood through observation and studies to develop a model that represents different kinds of "average users."

If someone asks what are the chances of finding evidence that this file was on this computer, we must be able to do better than answering "it depends."