Lauri Lalli

RISKS AND COSTS OF ERASURE FAILURES AND METHODS TO VERIFY THE ERASURE COST EFFECTIVELY

Statistics thesis
Department of statistics
Joensuu university
2008

ABSTRACT

As computers have become more widespread the amount of data on the hard disks has increased rapidly. When a computer becomes obsolete to its owner it is usually recycled. This can mean selling it to a new owner or melting the parts in order to get recycled materials. If the computer is sold to someone the hard disks may contains sensitive data if they are not properly sanitized.

Hard disks can be sanitized with suitable software. However, the sanitizing may not be fully effective if the hard disk is faulty or if the software is not suitable. Most hard disk sanitizing software contains some form of erasure verification. This thesis tries to find the optimal verification levels for different levels of data security.

In the case the hard disk contains very high security data it is reasonable to verify the hard disk by reading all its data. This is feasible when the customer doing the erasure feels that the highest level of security is required. If the data on the hard disk is of moderate or low security, the verification should read the 100 first and the 100 last sectors of the hard disk, and every 200th sector in between. This kind of erasure verification makes sure that at least most of the data on the hard disk has been erased.

TABLE OF CONTENTS

1 INTRODUCTION

As computers have become more widespread in our society, the amount of data gathered in their hard disks has also increased. While this may be a good thing as it makes a developed information society possible, it also leads to the risk of sensitive data leaks in a scale that was not possible at the time when the most advanced method of storing information was on paper in a file cabinet. Especially when these computers are recycled to new owners the data on the hard disks goes easily to unwanted hands unless the disks are properly sanitized.

Several studies have been made about the data that can be gathered from the hard disks of used computers. An especially noteworthy example is Craig Valli's "Throwing out the Enterprise with the Hard Disk" (Valli, C. 2004). Aquisti, A, Friedman, A. & Telang, R. have studied the cost of privacy breaches in their 2006 study. However, this thesis concentrates on smaller sub-problem: optimizing the verification of hard disk erasure. To this date no scientific studies exist on this subject.

When a hard disk is erased with some suitable software, the erasure can be verified by reading what data exists on the hard disk surface. However, reading through the whole hard disk takes a long time, so it is often not a feasible choice. This thesis tries to find an optimized method for hard disk verification in various situations. Doing this requires the use of statistical methods based on the data of older studies and knowledge of real-world challenges on the matter.

The aim of this study is to outline an optimal erasure verification scheme for hard disk erasure. This will be useful for any software company that produces erasure software, as it lays a scientific basis for the software implementation.

If the erasure verification finds even one sector that is not properly erased, then the erasure process has failed. In that case the person doing the erasure usually takes special care to destroy that hard disk by melting or pulverizing it. Any hard disk with data remaining after the erasure is a big liability to the company that owns it. On the other hand, when erasure verification has made sure that the hard disk is empty destroying it is not necessary, and the hard disk can be reused or sold. This is also better to the environment, as pulverizing or melting hard disks which are in perfect condition creates unnecessary waste.

2 COMPUTER LIFECYCLE

A computer is a device that processes information. The computer's hard disks store the information when the computer is shut off, unlike the device's main memory that is cleared every time the machine is powered off. Hard disks can store information for a long time. Many hard disks manufactured ten years ago are still in perfect working condition and the data they contain can be read. After the computer becomes obsolete its hard disks still retain data unless they are sanitized with suitable software or destroyed mechanically.

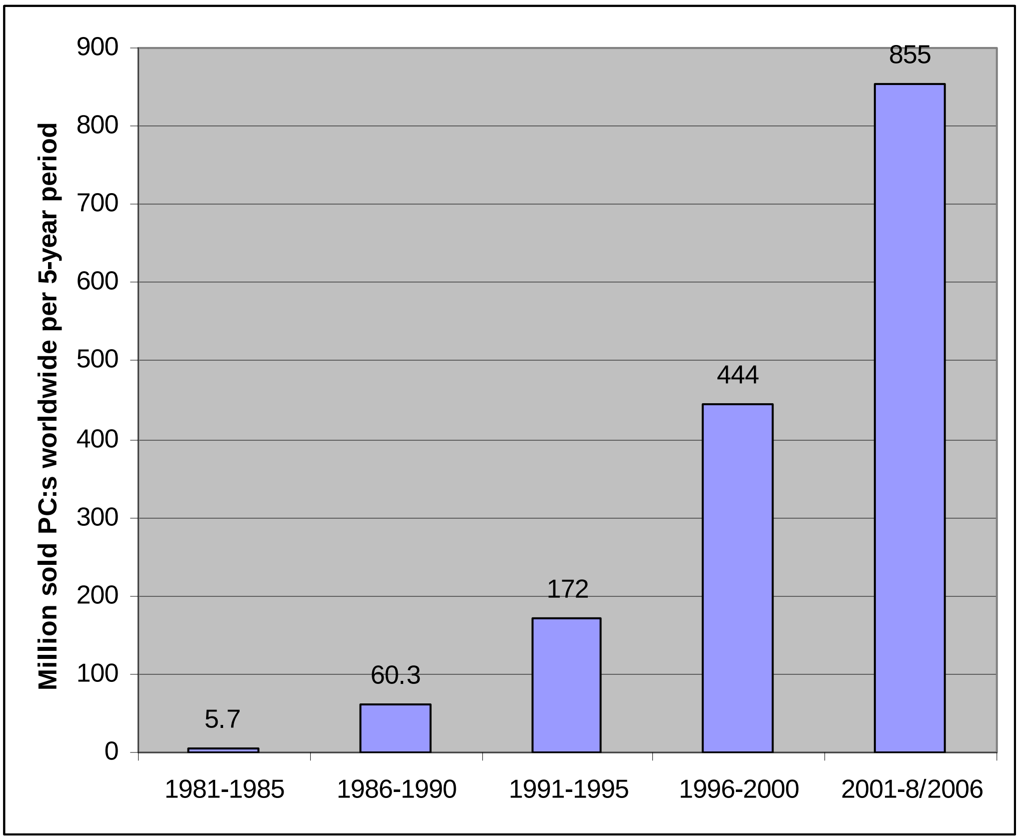


Image 2.1: Personal computer sales worldwide (Computer Industry Almanac 2006)

Computer sales are rising nearly every year. In 1983, when Time magazine declared the computer its "Machine of the Year", 2.8 million personal computers were sold worldwide. In 2005, sales surpassed 208 million units per year. (Grant, A., Meadows, J., 2006, 157) Another source, Computer Industry Almanac 2006 outlines the exponential growth of PC computer sales. The amount of units sold seems to double every 5-year period (picture 2.1).

Every computer manufactured has an economic lifespan before it becomes too slow for new applications. Matthews et al. assumes (1997, 4) that computers generally become obsolete to the purchaser in five years. After that there are four options available for the owner of the computer (Matthews et al. 1997, 3):

- the computer can be reused by the original owner
- the computer can be stored by the original owner
- the computer can be recycled
- the computer can be sent to a landfill

If a computer is recycled it can either be sold to a new owner or it can be melted down to get the precious metals the device contains. In the case the computer is resold in a working condition to a new owner, the new owner also gets any data the computer's hard disks hold.

Image 2.2: Personal computers waiting to be recycled (Nicholls, S, Kushin, M 2006)

Craig Valli (2004, 3) bought 11 used computers for his study and tried to recover data from the hard disks. Data was easily recovered in 10 cases out of 11. Most of these hard drives were either formatted or nothing had been done to prevent a data leak to the next owner of the computer. Valli describes that many of these hard disks contain confidential information that could harm the previous user if it found its way to the wrong hands.

3 SENSITIVITY OF THE DATA IN A HARD DRIVE

Computer users can be divided into three main subgroups: government, corporate and private users. The amount and type of data contained in the computers varies from group to group.

3.1 GOVERNMENTAL DATA SECURITY CLASSES

Governments are usually adept at defining data secrecy classes and methods for sanitizing computers. United States of America's National Institute of Standards and Technology has defined the sensitivity classes for governmental data (CMS Information Security Levels 2002, 1). The paper divides data sensitivity to low, moderate or high depending on what kind of data it is.

According to the paper (CMS Information Security Levels 2002, 1) high security data is the kind which when leaked to the wrong hands would cause catastrophic consequences. These might be, for example, some governmental arm losing mission capability for an extended period of time or the loss of human life.

Examples of governmental high security data are law enforcement records, intelligence agency information and any mission critical data which when leaked might cause loss of human life or other major consequences.

Leaking of moderate security data would cause serious consequences, for example major financial loss. For example, personal information about people working for the government or information about the internal workings of governmental bodies belong to this category. Passwords used on less mission critical systems are also moderate security data. Governments may also possess secret technologies that must be protected. Data about these is also of moderate security level. (CMS Information Security Levels 2002, 1)

Low security data is the least sensitive information. Generally any information that is readily available belongs to this category. This kind of information can be obtained for example from governmental internet pages. If this kind of data is leaked it may cause some minor consequences such as damage to someone's reputation. (CMS Information Security Levels 2002, 1)

3.2 CORPORATE DATA SECURITY CLASSES

Corporations do not have a universal method for assessing data security class. Different corporations have varying methods. Security guidelines are usually trade secrets and hence not readily available. Data on corporate computers can be arbitrarily divided into the same classes that the government of the United States of America uses: low, moderate and high sensitivity data.

High sensitivity data on a corporate computer would be something that would cause serious damage to the corporation if it got into the wrong hands. That kind of data might be, for example, secret plans for the corporation's future products or other trade secrets. If this kind of data find their way to competing corporations it might cause high monetary losses or even endanger the operation of the company.

Moderate sensitivity data on a corporate computer is anything that, if leaked outside the company, would damage the company or its customers but is not vital to the operation of the company. For example, in England a laptop was stolen containing the names and personal information of 11 million customers of the Nationwide building society (BBC News, Nationwide laptop theft 2006). This theft gained high-profile media publicity, because so many people's personal data was stolen. The data could be used for identity theft. Still, the theft did not endanger the functioning of the company, so the data was not high sensitivity.

Low sensitivity corporate data is anything that is not kept secret or can be found easily from the corporation's web site, from brochures or from any similar data source.

3.3 PRIVATE USER DATA SECURITY CLASSES

Private users' data can be divided into classes similar to governmental data: low, moderate and high sensitivity. Any data on a person's computer can be best classified by that person himself, as the sensitivity of the data in this case is mostly subjective. Old love letters might be highly sensitive from that single person's point of view, though they would not probably cause economic damage or loss of human life if they find their way to the wrong person.

High sensitivity private data would be anything that would cause serious damage to a person's life if made public or if the data was made available to the wrong person. For example, a person's computer might contain digital nude photographs of the owner of the computer. If these were made public it might cause serious stress to the person. In that case the damage is done to a single person, which is the main difference when compared to data theft from a company or a governmental computer, which might harm vast amounts of people. A single person’s suffering is still important to that person, though it might not cause damage to the whole society.

Moderate sensitivity personal data is anything that can cause economic damage or personal suffering if it is made public, but is not as damaging as high sensitivity data. This kind of data might be, for example, credit card numbers. If these numbers are stolen it can cause monetary losses to the person who owns the credit card, but usually credit card companies are willing to return the money to the victim of identity theft, so the loss is not catastrophic.

Low sensitivity data is anything about the person in question that is readily available, for example his favorite food or name. While this kind of information is sometimes best kept secret it will not harm the person if someone finds it out.

4 VALID METHODS OF SANITIZING HARD DISKS

Most standards for handling and erasing classified data come from governmental orders. United States of America’s Department of Defense has a program for assuring data security in governmental and industrial environments. This “National Industrial Security Program” (NISP) issued the Department of Defense directive 5220.22 on September 27, 2004. The directive outlines how sensitive data should be handled and sanitized. (DoD 5220.22, 2004, 1)

An attachment to DoD 5220.22, DSS Cleaning and Sanitation Matrix 2005, lists authorized methods of sanitizing hard disks and other data storage media. According to this paper, hard disks can be sanitized with a degausser or destroyed by disintegrating, incinerating, pulverizing or melting the hard disk. It is also acceptable to overwrite the hard disk with a character, then with the character complement, then verify the erasure’s success and lastly to overwrite the hard disk with random characters. It is also possible to verify overwrite success after the last random character overwrite, though that is in practice harder than verifying that the whole disk contains a known single character.

Valli’s study states that DoD 5220.22 “– – is used as a de facto compliance standard by manufacturers of erasure software” (Valli, C. 2004, 2). Almost all erasure software currently available is capable of using this erasure standard. It is widely accepted as a safe method of erasing a hard disk. The Defense Security Service guidelines still state that it is not good enough for erasing top secret information (DSS Cleaning and Sanitation Matrix 2005, 1). Top secret information could be safely destroyed for example by melting the whole hard disk in a very high temperature.

The English HMG Infosec Number 5 standard states that purging “restricted” data requires software that uses the HMG Infosec Number 5 Baseline standard erasure method. Erasing “confidential” or “secret” data requires using HMG Infosec Number 5 Enhanced standard erasure. HMG Baseline erasure requires overwriting the whole disk with a single character. For this standard erasure verification is not mandatory. The HMG Enhanced standard requires overwriting the whole disk with a character, then its complement and finally with a random character. For this method verifying the erasure is mandatory. These English standards require a similar erasure practice as DoD 5220.22. They are widely used, especially in the United Kingdom.

Peter Gutmann wrote a paper in 1996 stating that data overwritten once or twice may be recovered by subtracting what is expected to be read from a storage location from what is actually read (Gutmann, P, 1996). He also claimed that the data could be recovered even after many more erasure rounds, so all widely used methods of sanitizing hard disks are ineffective.

To overcome the fact that it may be possible to recover data from a hard disk even after it is overwritten Gutmann developed an erasure method that requires overwriting the hard disk 35 times: first 4 times randomly, then 27 times with specific patterns and last 4 times randomly (Gutmann, P 1996). While this may be a very secure way of overwriting a hard disk it is such a slow method that it is not practical. Overwriting an average 200 GB hard disk just one time takes maybe an hour. Overwriting it 35 times would take 35 hours. This is far too long for any real world application.

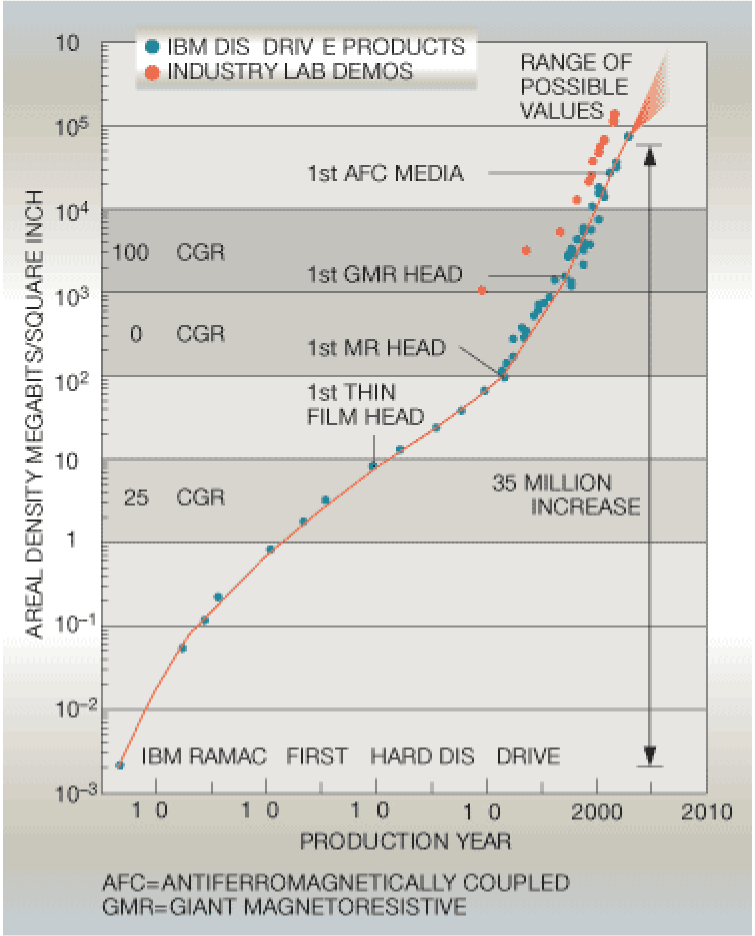


Image 4.1: Development of hard disk areal density (Grochowski, E., Halem, R. 2003)

Recovering data after overwriting has become harder and harder as hard disk magnetic densities have increased. Valli states that “– – recovery by magnetic remnant imaging is fast becoming an infeasible attack. (Valli, C. 2004, 2)” The magnetic density of a modern hard disk is so high that it is probably not possible to recover overwritten data by trying to read the area between the hard disk’s magnetic tracks using magnetic microscopy. Many methods that were feasible in the year 1996 when Gutmann wrote his paper are not usable anymore. If a hard disk is overwritten three times according to the DoD 5220.22 standard, probably not even the secret services of large countries are able to recover the data.

Disk sanitizing guidelines such as the Department of Defense directive 5220.22 and HMG Infosec Number 5 standard give a solid basis for developing a unified disk sanitating guideline. Disk sanitizing strength should be based on the sensitivity of the data contained in the disks. The data can be divided to high sensitivity, moderate sensitivity and low sensitivity. This division is valid independent of where the disk comes from, though decision criteria for placing the data in the different classes may vary according to whether the hard disk was in governmental, industrial or private use.

Data security class	Sanitizing method
High	Overwrite the hard disk once with a character, then with the character complement, verify and overwrite the last time with random characters. After that physically destroy the hard disk by shredding it to tiny pieces. Notice that erasing the hard disk before physical destruction is necessary to make sure that no data can be recovered from the pieces of the hard disk by magnetic microscopy or other high-tech methods. Physical destruction is approved for sanitizing top secret data (DSS Cleaning & Sanitation Matrix 2005, 1)
Moderate	Overwrite the hard disk once with a character, then with the character complement, verify and overwrite the last time with random characters (DSS Cleaning & Sanitation Matrix 2005, 1). This sanitizing method is approved by the Department of Defense for erasing all except top secret data.
Low	Overwrite the hard disk once with zeroes and verify the erasure. This method is outlined in the HMG Baseline Standard Number 5 (HMG Infosec Number 5, 2003, 13)

Table 4.1: Erasure methods for different data sensitivity classes

5 REASONS FOR ERASURE FAILURES

Most hard disks sold after use are not erased. This is confirmed by Valli’s research, as 10 out of 11 hard disks he purchased and studied were not erased properly (Valli, C. 2004, 3). A major reason for erasure failures is that no one has tried to sanitize the hard disk. If the user just formats the hard drive or does nothing to protect the data on the hard drives, nothing stops the data from getting to the wrong hands.

In some cases the users think that their computers are properly sanitized before they are sold although they are not. Some recyclers ship the old computers to third world countries, and do not necessarily sanitize the hard disks before the shipping. For example Yemen does not have e-waste laws (Boran, O. 2004, 18). Foreign companies ship used computers to Yemen, where they are usually sold in small workshops. If a company’s computers end there without proper sanitizing there is a risk of misuse of the data contained in the computers.

Sometimes a computer breaks down and has to be replaced. In some cases the user thinks that the hard disk of the broken computer is not working, and thus needs not be sanitized. However, it is plausible that a computer may break down so that the hard disk is still intact. If the hard disk is transferred to a working computer the data may be recovered.

In the case the computer is sanitized before it is sold, the sanitizing method may be faulty. For example, formatting a hard disk is not enough to erase data from the hard disks. If the hard disk is overwritten, it may contain a “host protected area” that may not be overwritten if the program used for sanitizing the hard disks does not support overwriting host protected areas. It is also possible that the program does not detect hard disk size correctly, and thus erases only part of the data.

If the hard disk to be sanitized is mechanically or electronically broken, it is possible that data on the disk can still be read but it cannot be correctly overwritten. In this case sanitizing the hard disk by means of software fails and the hard disk should be mechanically destroyed by pulverizing or melting it. It is also possible that the software used for sanitizing the hard disk is not overwriting the data correctly for some unspecified reason. These erasure failures may lead to sensitive data leaking to whoever buys the computer when it is recycled to the next owner.

In the case the hard disk is sanitized using suitable software, there are some possible scenarios where the whole hard disk surface is not sanitized. The types of these errors can be divided into two main classes: the hard disk is detected incorrectly or the overwriting of the data is done incorrectly.

5.1 HARD DISK IS DETECTED INCORRECTLY

There are several reasons why the hard disk can be detected incorrectly. Most often this happens when the computer's BIOS detects the hard disk size incorrectly and the erasure software relies on BIOS information. For example, the BIOS reports that the hard disk size is 137 GB even though the real hard disk size is 250 GB, because of the ATA interface limit barrier. If the hard disk contains data in the area above 137 GB limit, that is not erased if the erasure software relies on BIOS information. This is not the only known hard disk BIOS barrier; there are many others that limit hard disk detection. See table 5.1.1 for some well-known hard disk size barriers. These hard disk size barriers affect mostly older computers made before the year 2000.

Hard Disk Size Barrier	Size	Notes
PC/XT Parameter	10.9 MB	Design limitation of first PC/XT computers
FAT12 Partition Size	16.7 MB	4086 clusters of 4096 bytes
DOS3 Barrier	33.6 MB	16384 clusters of 2048 bytes
DOS4 Barrier	134 MB	65526 clusters of 2048 bytes
Standard IDE/ATA	528 MB	Max cylinders 1024, max heads 16, max sectors 63
4095 Cylinder Limitation	2.11 GB	Enhanced IDE/ATA BIOS
FAT16 Partition Size	2.15 GB	File system limitation
6322 Cylinder Limitation	3.26 GB	Rare limitation
Phoenix Bios Bug	3.28 GB	Rare limitation
8192 Cylinder Limitation	4.22 GB	BIOS geometry translation limitation
240 Head Int13 Interface	7.93 GB	BIOS structure limit
Int13 Interface Barrier	8.46 GB	BIOS structure limit
Windows 95 Limit	32.0 GB	Operating system limit
65536 Cylinder Barrier	33.8 GB	BIOS geometry translation limitation
ATA Interface Limit	137 GB	Upper limit of ATA standard

Table 5.1.1: Well-known limits to hard disk size (Storagereview.com 2005).

A hard disk can contain a Host Protected Area (HPA) or a Device Configuration Overlay (DCO). These are both areas of the hard disk that cannot easily be accessed by normal programs. As Gupta, Hoeschele & Rogers claim in their 2006 paper, “These areas can be problematic for computer forensics investigators, since many of the common industry tools cannot detect the presence of HPA and DCO”. These areas can be problematic to any company manufacturing hard disk sanitation software, since it is a non-trivial task to dismantle the HPA or DCO so that their contents can be erased. For this reason it is possible to detect hard disk size incorrectly if the hard disk contains HPA or DCO. In that case not all data on the hard disk is erased properly by the erasure software.

Hard disks can be partitioned to various sector sizes. The most common sector size is 512 bytes per sector, but other sizes exist. The most often used other sector size is 520 bytes per sector. It is possible that the erasure software detects only the first 512 bytes of a given sector if the real sector size is 520. This error can cause some data to remain after erasure.

5.2 OVERWRITING DATA IS DONE INCORRECTLY

If the hard disk is detected correctly, then in an ideal situation all the data is erased correctly. However, if the hard disk that is being erased is not working properly the erasure can fail. The most common situation is that the hard disk that is being erased contains so many bad sectors that the erasure time is very long, and after the erasure is performed some data can remain in the bad sectors that could not be overwritten.

It is possible that the tracking of the hard disk writing head fails when overwriting the hard disk. This may happen for example if the hard disk is somewhat broken and overwriting some area of the disk takes a very long time, which may cause the overwriting software to skip to the next area. In that case some areas that still contain data can remain.

The third possible reason for the hard disk not being properly overwritten is a software error in the erasure software. As computer programs are becoming increasingly complex it is not possible to verify that any given software performs perfectly in all possible scenarios. The chance of this error can be lowered by testing the erasure software thoroughly, but the possibility of an error always remains.

6 METHODS FOR VERIFYING HARD DISK ERASURE

The safest method for verifying hard disk erasure is reading through the whole hard disk surface. If the hard disk is mechanically fully functional all the data on the disk can be read and thus verified. This is the upper limit of software-based verification. No other method can uncover data that reading all the data on the hard disk did not reveal. Of course the hard disk may be mechanically faulty so that both overwriting data and verifying the overwriting fail, but if that is the case then software-based hard disk sanitizing may not be the best alternative for data destruction. Mechanically pulverizing the hard drive or melting it is a safer way to destroy broken hard drives.

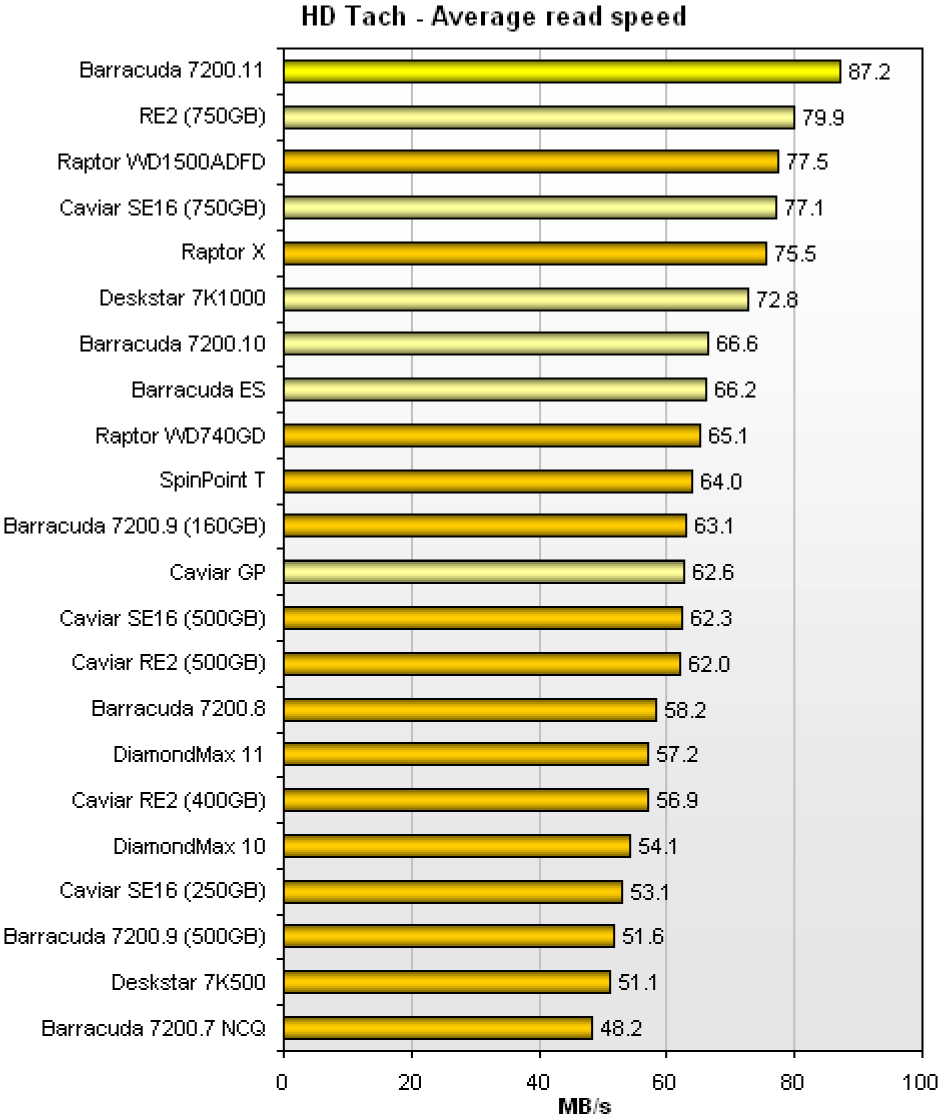


Image 6.1: Average hard disk read speed (Geoff, G. 2007, 13)

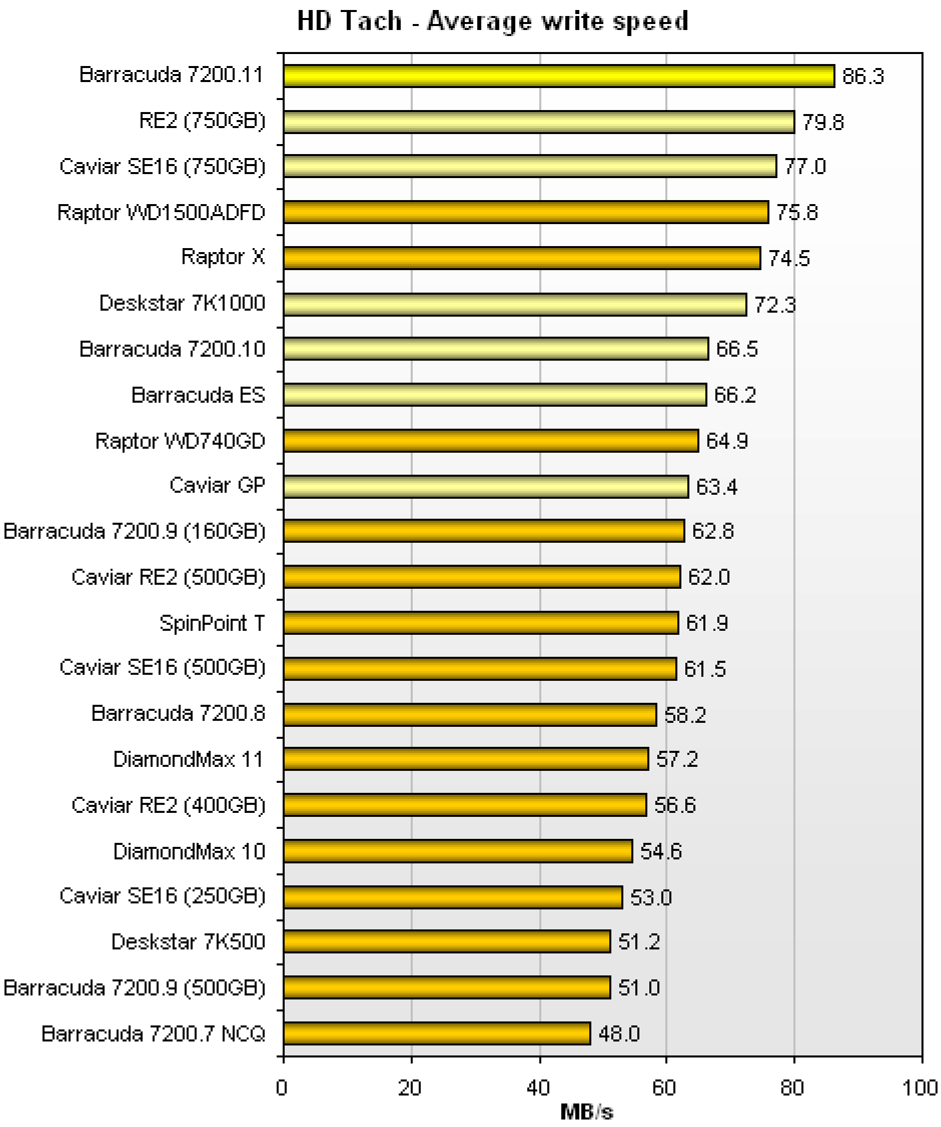


Image 6.2: Average hard disk write speed (Geoff, G. 2007, 13)

Reading all the data on a hard disk takes about the same time as writing the hard disk full of data. See pictures 6.1 and 6.2 to verify this. This means that if every hard disk erasure is verified by reading the whole hard disk, the erasure time composed of erasing and verifying doubles compared to only erasing the hard disk. This is most often not economically feasible.

Statistical methods of sampling allow calculating the probability of the whole disk being erased even if the whole disk is not read after the erasure. Most often these methods involve taking a small number of samples from the disk surface, for example ten samples where each sample covers one sector of the hard disk. If all the data on these samples is erased, then it is feasible to think that the erasure has been a success. The certainty of this knowledge can be calculated.

Verifying the erasure also faces the problem of detecting the difference between erased and unerased data. All the bits on the hard disk surface are either zeroes or ones. If the last erasure round before verifying was zeroes, all the data on the hard disk is likely to be zeroes. However, if the hard disk that was erased contained only a little data, then it is possible that almost all the bits are zeroes even if no data is erased.

If the last overwriting round before verification uses real random overwriting, then data on the disk is not just zeroes but random zeroes and ones. In this case detecting the difference between erased and unerased data is probably impossible. Also it is useful to notice that widely used programs for viewing the raw data on hard disk surface do not show bits, they show bytes. Thus visible data on the hard disk may be shown as many other figures and letters besides zeroes and ones.

One important consideration about erasure verification is that verification should be independent of erasure. If, for example, both the erasure part and the verification part of the software use common code for determining the hard disk size, it is possible that both the erasure software and the verification software give the same wrong result. In that case the verification is useless.

In practice both erasure and verification are integrated into the same software, so they have to share some code and at least have a common user interface. Still it is a good idea to try to make these parts of the software as independent as possible.

Different erasure failures can be found by different kind of verification systems. The most effective method is to read through the whole hard disk. If the hard disk size detection is done right in the verification software, and the software is able to see the contents of Host Protected Areas (HPA) and Device Configuration Overlays (DCO), this method would uncover all erasure errors. However, this is usually not a practical way of verifying the erasure.

Generally, if the error in the erasure process has been incorrect detection of the hard disk size, a single sample from the area that was not erased usually shows that the erasure was done incorrectly. The problem is to get the verification software to detect the hard disk size even more reliably than the erasure software, which can be problematic. However, if we consider a perfect verification software, a very low number of samples would show that the erasure was done incorrectly.

In the case the erasure has failed because of bad sectors on the disk, hard disk tracking errors or an erasure software bug, it is probable that most of the hard disk is erased but random unerased areas remain on the hard disk surface. In the case of bad sectors these would be the size of sectors, most often 512 bytes. Bad sectors can be clustered together, for example if the reason for the bad sectors is a head crash where the writing heads of the hard disk have collided with the hard disk surface. Most often, though, they are dispersed around the hard disk surface.

The problem now is: “how many sectors s must be read in order to verify that there are less than $i_{\max}$ unerased sectors, where i is the number of unerased sectors on the hard disk and $i_{\max}$ is set by the person defining the requested safety level of the erasure”.

Any kind of flaw in the erasure process is likely to produce a number of unerased sectors. If these are detected, the erasure software can inform the operator of the software that the erasure failed and the hard disk must be disposed of by other means, most likely by mechanically breaking it into small pieces or by melting it in a very high temperature.

How dangerous these unerased sectors are depends on the contents of the hard disk, and in a way on the sensitivity class of the data. For example, if the hard disk is filled with credit card numbers and the personal information of a very large amount of people, it is very likely that any unerased sector may contain some sensitive information. If, on the other hand, the hard disk that has been erased belongs to a home computer that does not have much sensitive data, it is very unlikely that any given sector contains any damaging information.

The verification method can be developed based on the density of the sensitive data on the hard disk; if sensitive data is very rare and the amount of unerased sectors is low it is very likely that the hard disk does not contain any sensitive information after the erasure even if the verification is not able to detect that the erasure has failed.

7 HARD DISK DATA STRUCTURE

Let’s assume the computer has only one hard disk, and that hard disk is new and empty when a person obtains the computer. The first things he or she does are to format the hard disk and to install the operating system. Let’s assume the person formats the whole hard disk into a single partition. Now, when he or she installs the operating system it will be at the beginning of the hard disk, that is, closest to the outer edge of the hard disk. The hard disk will generally fill from the beginning to the end. Only if data is erased from the hard disk, gaps of empty space appear in the middle of the data.

Let’s say the hard disk size is 100 GB and Windows XP is installed to it. The operating system occupies the first 4 GB in the beginning of the hard disk, and the paging file about 2 GB more. At this point the rest of the hard disk is empty. The user will probably install the programs he or she frequently uses next. This takes probably around 2 GB more on an average computer. At this point the user has put about 8 GB of data into the hard disk. These are at the beginning of the hard disk, and at this point nothing really personal is on the computer.

Next the user probably transfers his personal data to the computer. This may take any amount of space, but probably much less than the hard disk’s maximum size. The reason for this is that if the user had more data to store, he or she would have bought a bigger hard disk. At this point the hard disk contents are as follows:

Area Size	Area Begins	Area Ends	Contents
8 GB	0 GB	8 GB	System Files
X GB	8 GB	8 + X GB	Personal Files
100 – (8 + X) GB	8 + X GB	100 GB	Empty Space

Table 7.1: Hard Disk Contents

Although the whole hard disk should be erased, the most important area to be erased lies between 8 gigabytes and 8 + X gigabytes, where X is the amount of files the user has put into the hard disk. A possible system for verifying the erasure while taking this knowledge in account would be taking two sets of samples from the hard disk surface. Set 1 would contain points from the whole hard disk, and set 2 would contain points that lie between 8 GB and the last data found on the hard disk. The last point on the hard disk that contains data can be found from the hard disk's file index. The second set of verification samples can thus be taken from an area beginning some gigabytes from the beginning of the hard disk, and the samples can end at the point where the last data was detected.

This assumption about the contents of the hard disk helps in developing the verification scheme, but it relies on theoretical use of the hard disk. If the hard disk is in use for a long time and data is deleted and overwritten many times it is possible that a piece of sensitive data can lie almost anywhere on the hard disk. Even in that case the density of the sensitive data would usually be greater near the beginning of the user data, just after the operating system files.

It is possible to divide the sectors on a hard disk into two categories: those that contain sensitive data and those that do not. If the computer to be erased belongs to an average person and the general sensitivity of the data is low, it is probable that the whole computer has only a couple of sectors containing sensitive data. These pieces of sensitive data might be the social security number and credit card number of the user of the computer and maybe the Windows authentication code. If that happens to be the case, there would be three sectors containing some sensitive data. If the hard disk size is 100 GB it contains 195358867 sectors, so about one sector in 65 million contains sensitive data.

In the case the hard disk belongs to a big company that stores one million customer records in the hard disk, the computer data can be considered highly sensitive because if those customer records find their way to the wrong hands they can be very damaging to the company. In that case about one sector in 200 contains sensitive data. The chance that sensitive data can be found from areas that are not properly erased is thus much higher.

When developing optimal verification schemes, the density of sensitive data on the hard disk must be taken into consideration. The verification of erasure can give a conservative estimate of how big a part of the hard disk is erased, and thus it is possible to assess how large an area might remain unerased. Using probability calculations it is then possible to calculate how probable it is that the hard disk still contains some sensitive data after the erasure, considering that the erasure process may be flawed and leave some sectors unerased because of various reasons, such as bad sectors.

8 COST OF VERIFYING A HARD DISK ERASURE

Erasure verification increases the time required for the whole erasure process. The erasure process costs more to the client if it takes longer, because someone performing the erasure either gets paid for performing the erasure or someone loses working time that could be used for other purposes. If the erasure takes only a small amount of time, which is the case when only a limited amount of samples are taken from the hard disk, the cost involved is minuscule. However, if the erasure verification takes a long time the cost is high.

If a recycling center is doing the erasure, they get paid a set amount for performing the task. The United States Census Bureau states that an average manual laborer in United States in the year 2005 got paid 11 dollars and 43 cents per hour, if he or she worked in the private industry (U. S. Census Bureau table 624, 2008). In addition to the salary of the worker come other expenses such as the cost of the building where the computer recycling takes place. These other costs vary from recycler to recycler. On the other hand a single worker can usually perform the erasure of the hard disk on several computers at the same time, which lowers the cost per computer. Considering that the company doing the erasure has other costs besides the salary of the worker, but the worker can erase multiple computers at the same time, a rough estimate would be that every additional hour of erasing a single computer costs $10 to the recycling company.

If the erasure verification is performed by reading every point of the hard disk surface, the time required would be about an hour for an average 100 GB hard disk that is being recycled. Thus, this erasure verification model would increase the cost of recycling the computer by $10, which is a major cost increase when recycling the computers of average users. Erasure verification by taking samples from the hard disk would be much faster, and thus make recycling the computer cheaper.

Some clients such as banks do not care how much time the erasure takes. They are more concerned about the data being thoroughly sanitized. This difference in thinking can be seen also in the erasure method selection. Most computer recycling centers use an erasure method that overwrites the hard disk once with zeroes. This is enough to prevent almost all current data recovery methods. Banks use most often either the DoD 5220.22 standard or some other standard that overwrites the hard disk several times. For these clients the time cost of erasure is not an issue.

9 COST OF ERASURE FAILURES

There are two main kinds of cost involved with security breaches. One is easier to measure and more tangible: the direct monetary costs of a security breach. These might involve notifying the customers whose identity has been stolen, fines laid to the company in a court and other costs of recovering from a security breach. For example, the company Chokepoint involuntarily released 163000 consumer credit reports in the year 2005 and was forced to pay $15 million in penalties, which is about 10% of its $143 million earnings in the year 2005. (Aquisti, A, Friedman, A. & Telang, R. 2006, 2).

A second type of cost involved with a security breach is damage to the brand of the company. This is not as easy to measure, but can be inferred from changes in the company's stock price following the security breach. Cavusoglu, Mishra and Raghunathan considered this in their 2004 paper that was published in the International Journal of Electronic Commerce. They notified that following the publication of a security breach a company's stock lost on average 2.1 percent of its market value. This means on average $1.65 billion market value loss per company security breach.

This kind of cost to the company is realized when a security breach really happens. If a hard disk is not thoroughly erased before it is sold to the next user, it is entirely possible that nothing bad happens. The next user overwrites the hard disk with his or her own data and does not even realize that he or she has some sensitive data on the hard disk of his or her computer. However, every time a hard disk is not erased properly there is a possibility of economic costs. Verifying the erasure should give the company which is selling the old computer certainty that there is no harmful data left on the hard disk.

The cost of a security breach also involves other users of computers besides companies. In the case of governments there might be diplomatic consequences if secret data is found on a hard disk that is decommissioned. If the computer belongs to an average person, the monetary cost of the security breach is not as great as in the case of companies, which have the data of large group of people on their computers, but the data leak may cause embarrassment and suffering nevertheless.

It seems that the average cost of a security breach depends on two things: the amount of sensitive data on the computer and the sensitivity of data on the computer. The biggest recorded losses to companies have been cases when large amounts of sensitive data have found their way to the wrong hands. The 2006 study by Aquisti, Friedman and Telang found that a median breach disclosed the data of 95000 subjects. The same study found that a breach of more than 100000 subjects will reduce the stock price of the company by an average 1.2% (p=0.077). This is a smaller figure than the one given in the 2004 paper by Cavusoglu, Mishra and Raghunathan, which claimed that an average security breach causes 2.1% loss to the stock value of a company. The damage to a company caused by a single item of information released to the wrong hands by a security breach is a bit arbitrary and defining it is outside the scope of this thesis. However, some kind of estimate is necessary as a basis for the following calculations. The companies studied by Cavusoglu, Mishra and Raghunathan had an average total stock value of $78.5 billion, calculated from the information that 2.1 percent of the value is $1.65 billion. Based on 2006 study by Aquisti, Friedman and Telang a median breach compromises 95000 counts of sensitive data and causes 1.2% stock value loss, which is about $1 billion based on data by Cavusoglu, Mishra and Raghunathan. Stock prices recover in time, so this cost is not permanent. Direct monetary costs must be added to this number, but they are insignificant compared to the stock value loss.

The above calculation is based on information about companies. The price of unhappiness caused by a single person's sensitive data going to the wrong hands when he or she sells his or her old computer and does not properly sanitize the hard disk is even harder to quantify. If the computer contains, for example, nude photographs of the former owner of the computer and these photographs find their way to the internet, the damage is mainly the shame and unhappiness of the former owner of the computer. Setting a price to unhappiness is not easy and it is not within the scope of this thesis. The figure calculated for the companies can be used as an estimate for the loss experienced by a single person. Some people feel that they are willing to compromise their reputation for $10000; some others feel that it is not enough money.

Computers used by governments containing secret data are security liabilities. If the data finds its way to the wrong hands the damage can be great. If compromised information damages the reputation of a country or compromises the mission capabilities of a governmental organization it is not easy to set a price to the damage. The figure calculated for the losses for companies can be used as an estimate for the losses for governments, but this is just a rough estimate.

When a large amount of customer data has leaked from a company, the hard disk that contains the information has probably not been erased. If the hard disk is at least partially erased the amount of remaining sensitive data is lower, and thus damage caused by it is not as great. Still even a single piece of sensitive data in the wrong hands can harm the former owner of the hard disk.

10 MATHEMATICAL MODEL OF THE HARD DISK

A hard disk is a device with spinning platters that contain data.

Image 10.1: Hard disk (University of Utah)

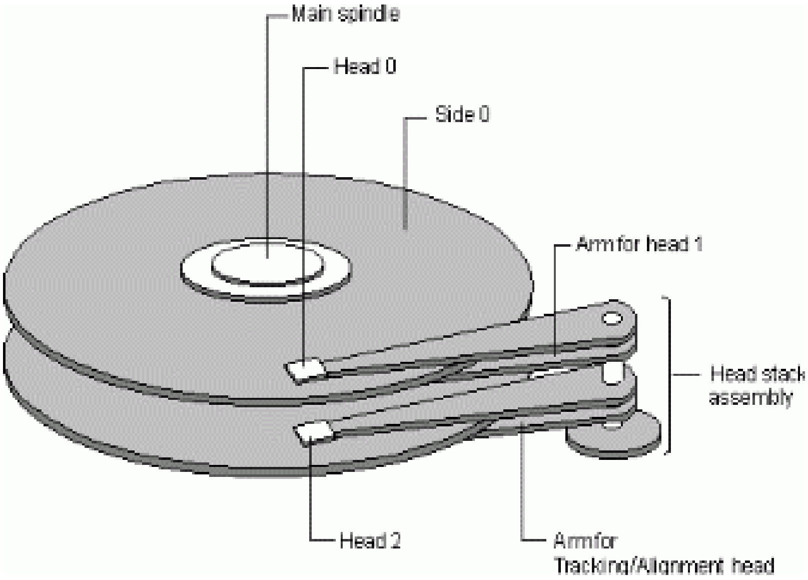


Image 10.2: Hard disk structure (Windows NT Server Resource Guide)

Magnetic heads at the end of arms read and write data to the platters. The data is written to tracks on the platters. The tracks are divided into sectors. Near the edge of the hard disk a track has more sectors than near the center of the hard disk.

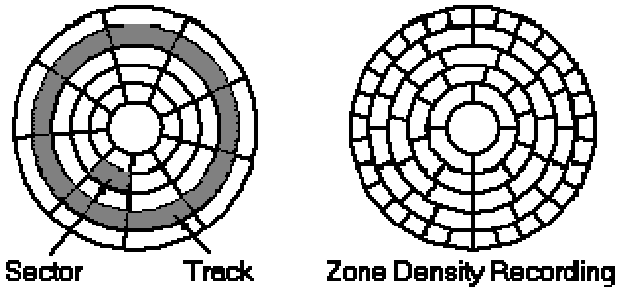
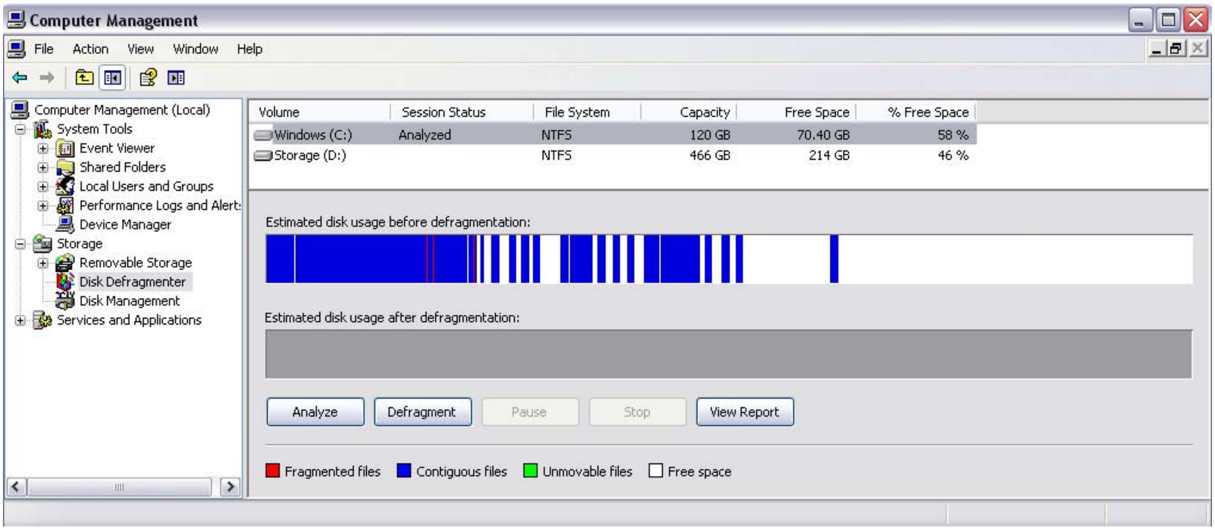


Image 10.3: Hard disk sectors and tracks (Dew Associates Corporation Knowledge Center)

The sectors can be arranged into a line starting from the beginning of the hard disk. This leads to a mathematical model of the hard disk: the data lies on a single line segment.



Volume	Session Status	File System	Capacity	Free Space	% Free Space
Windows (C:)	Analyzed	NTFS	120 GB	70.40 GB	58 %
Storage (D:)		NTFS	466 GB	214 GB	46 %

Estimated disk usage before defragmentation:

Estimated disk usage after defragmentation:

Image 10.4: Data on a computer hard disk as seen by Windows XP Disk Defragmenter. Image taken from the home computer of the writer of the thesis.

Windows XP Defragmenter is a Microsoft tool that shows how the contents of the hard disk are located, as image 10.4 shows.

A mathematical model for the hard disk could be a line segment that contains all sectors of the hard disk from the beginning to the end.

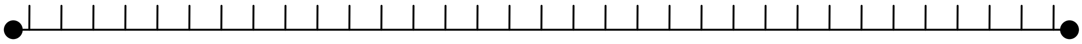


Image 10.5. Mathematical model of the hard disk. Smaller segments are single sectors.

Image 10.6. Mathematical model of a single sector. Smaller segments are bytes of data. Each sector contains 512 bytes of data.

Data on the hard disk can be compared to a long row of baskets, each holding either a zero or a one. This means that discrete models are in this case more precise than continuous ones.

11 MATHEMATICAL MODEL OF ERASURE ERRORS

Possible errors in the erasure process have been listed in chapter 5. It seems that the errors are of two subtypes: either large sections of the hard disk being not erased due to large errors such as detecting the hard disk size wrong, or small sections of the hard disk not being erased correctly due to remapped sectors or similar other problems. These two problems are very similar and differ mainly in the scale of the problem, but it is reasonable to consider these as two subproblems.

Definition 11.1: Large scale erasure errors are henceforth called type I erasure errors. These affect series of sectors on the hard disk and usually cover areas larger than 1 MB.

Definition 11.2: Small scale erasure errors are henceforth called type II erasure errors. These affect single sectors or single bytes inside sectors and usually cover areas smaller than 1 MB.

11.1 POISSON PROCESSES

Çinlar defines (1975, 71) that an arrival process $N = \{N_t; t \geq 0\}$ is called a Poisson process provided that the following axioms hold:

a) for almost all ω, each jump of $t \to N_t(\omega)$ is of unit magnitude;

b) for any $t, s \geq 0$, $N_{t+s} - N_t$ is independent of $\{N_u; u \leq t\}$;

c) for any $t, s \geq 0$, the distribution of $N_{t+s} - N_t$ is independent of t.

Type I erasure errors lie on the hard disk between arbitrary points. Because type I errors are on sector scale, covering single or more sectors, then part a) of Çinlar’s definition holds true. The process of arrival of error area is discrete with each jump $t \to N_t(\omega)$ of unit magnitude. Part b) of Çinlar’s definition states that $N_{t+s} - N_t$ is independent of what has happened before t. In case of type I errors this seems to be true. The final part of Çinlar’s definition states that the distribution of $N_{t+s} - N_t$ is not dependent on t. Because it is a likely premise that error rate remains the same in all points of the hard disk axiom c) holds also true. Erasure errors on the hard disk can be modeled as a Poisson process.

As Çinlar defines (1975, 76), $N = \{N_t; t \geq 0\}$ is a Poisson process with rate λ if and only if:

a) for almost all ω, each jump of $t \to N_t(\omega)$ is of unit magnitude;

b) for any $t, s \geq 0$, $E[N_{t+s} - N_t \mid N_u; u \leq t] = \lambda s$

This means that for a Poisson process with rate λ the history before t does not affect what happens after t, and expected value of the process in a given span of time from t to s is λs. Notice that λ is the rate of events, and can be derived from real-world observation.

Poisson process with rate λ has some interesting properties. As Çinlar notices (1975, 79), the process can be used as a model for the time of arrivals. For this process $t \to N_t(\omega)$ is non-decreasing, right continuous and increases by jumps of size one. The function is completely determined by the times of its jumps, let these be $T_1(\omega)$, $T_2(\omega)$, … so that T_1, T_2, … are successive instants of arrivals. Çinlar (1975, 79) proves that if t is an arbitrary point in time the probability that there are no arrivals in $(t, t+s]$ is $e^{-\lambda t}$, independent of the history of arrivals before t. This convenient property holds true also in case t happens to be the time of arrival T_n. This can be written down as:

$$P\{N_{t+s} - N_{T_n} = 0 \mid N_u; u \leq T_n\} = e^{-\lambda s} \qquad \text{(Formula 11.1.1)}$$

Here $\{N_u; u \leq T_n\}$ is the history of the process until time T_n, which is time of the nth arrival. As Çinlar (1975, 79) states, the information contained in the history $\{N_u; u \leq T_n\}$ is the same as that contained in $\{T_1, ..., T_n\}$, which can be restated as:

$$P\{T_{n+1} - T_n \leq t \mid T_0, ..., T_n\} = 1 - e^{-\lambda t}, t \geq 0 \qquad \text{(Formula 11.1.2)}$$

Here we have written $T_0 = 0$. So, arrival time T_1 and the following time periods $T_2 - T_1$, $T_3 - T_2$ and so on are independent and identically distributed random variables with the common distribution being $1 - e^{-\lambda t}, t \geq 0$. As this distribution is exponential distribution with parameter λ, it has probability density function of $\lambda e^{-\lambda t}, t \geq 0$.

Exponential distribution is memoryless (Çinlar 1975, 80), which can be written as:

$$P\{X > t + s \mid X > t\} = P\{X > s\} \quad t, s \geq 0 \qquad \text{(Formula 11.1.3)}$$

Notice that exponential distribution is the only one that has this property, so all distributions having this property are cases of exponential distribution. For exponential distribution the following hold true (Çinlar 1975, 81):

$$Var(T_{n+1} - T_n) = \frac{1}{\lambda^2} \qquad \text{(Formula 11.1.4)}$$

$$Var(T_n) = \frac{n}{\lambda^2} \qquad \text{(Formula 11.1.5)}$$

$$E[T_n] = \frac{n}{\lambda} \qquad \text{(Formula 11.1.6)}$$

11.2 MODEL OF LARGE SCALE ERASURE ERRORS ON HARD DISK

Some erasure failures generate large areas of hard disk that are left unerased. This kind of errors are categorized as type I erasure errors in definition 11.1. Reason for this kind of errors can be for example detecting the hard disk size wrong or some other inadequacy of the erasure software.

The erasure failure can be modeled as a Poisson process. There is a chance that the area of failed erasure starts on each of the sectors of the hard disk. This chance is determined by the parameter λ of the process. The value of λ depends on many things. If the hard disk is new and the software being used is suitable for erasing it the value of λ is probably so low that errors occur very rarely. If a hard disk is worn or partially broken the value of λ gets higher.

The value of λ is in most cases so low that this kind of erasure error occurs very rarely, maybe once every one thousand erased hard disks, again depending on the program used for sanitizing the hard disk. In this case arrival time T_1 marks the spot where the erasure error begins. The area affected by the failure can continue until the end of the hard disk, or in some cases it can be smaller area. To model this behavior, another Poisson process starts at T_1 with the parameter λ_2, where $\lambda_2 > \lambda$. The reason for this is that if λ_2 is as high as λ, it is very unlikely that the affected area of the hard disk ends before the end of the hard disk.

In case the hard disk contains a type I erasure error, the unerased area begins at T_1. The area ends when the Poisson process with parameter λ_2 generates an end-point to the area, or when the end of the hard disk is reached. If this endpoint is written as T_E, then the size of the unerased area is:

$$U_1 = T_E - T_1 \qquad \text{(Formula 11.2.1)}$$

11.3 MODEL OF SMALL-SCALE ERASURE ERRORS ON HARD DISK

Small scale erasure errors, or type II erasure errors, happen because of remapped sectors on the hard disk or because of other small-scale errors. Most hard disks do not contain this kind of errors, but those that have them probably contain many small patches of errors. Thus it is reasonable to divide hard disks into two classes: those that contain this kind of errors and those that do not. In the case of a hard disk that has these errors a Poisson process deposits the error areas randomly. Each time the Poisson process with parameter λ initiates an erroneous area, another Poisson process with parameter λ_2, where $\lambda_2 > \lambda$ starts. This arrival time process simulates how the erroneous area on the hard disk ends. It is very likely that parameter λ_2 is in this case such that the arrival of the end of erroneous area happens quickly after the area begins. It is very rare that there are more than a couple of consecutive remapped sectors, though some hard disk failures such as a head crash where the reading heads connect with the disk surface can create larger error areas.

In practice the amount of type II erasure errors varies, but the most likely scenario is that less than 0.1% of the hard disk sectors are damaged in a way that results in type II erasure errors during erasure. The average size of such erroneous areas is a bit over one sector, as the majority of remapped areas are exactly the size of one sector, but there are a few larger erroneous areas. It is also possible that some sectors are slightly broken so that less than one sector is not sanitized when the hard disk is erased. It seems that for the Poisson process that ends the type II erasure error area λ_2 should be selected so that the expected size of the erroneous area is about one sector.

Because $E[T_n] = \frac{n}{\lambda}$, the expected amount of type II erasure errors depends on λ and the disk size n. The amount of sectors that are left unerased because of type II erasure errors is $E[T_n]$ multiplied by the average size of the erroneous area. If the expected size of erroneous area in case of type II erasure errors is one sector, then the size of unerased area because of type II erasure errors is

$$U_2 = E[T_n] \qquad \text{(Formula 11.3.1)}$$

where U_2 is the number of sectors which have failed to be erased.

11.4 CATASTROPHIC ERASURE ERROR

In some rare cases the erasure software does not erase anything from the hard disk. In this case the whole hard disk from the beginning to end contains an unerased area. In that case, the size of the unerased area equals the size of the hard disk. This kind of catastrophic erasure error is rare, but they do happen occasionally with some combinations of hardware and software.

12 DETECTING ERASURE ERRORS ON HARD DISKS

As noted in chapter 11, there are three main types of erasure errors. Each of these creates a different kind of threat to information privacy, and detecting them demands three different approaches. Detecting a catastrophic erasure error is the easiest of the three, because it is usually easy to notice that no data has been erased. If a hard disk is not erased there is most probably data in the beginning of the hard disk, where for example boot sectors usually are. Reading the first 100 sectors of the hard disk should thus reveal this kind of error with high certainty.

In the case of catastrophic erasure error no data is erased, and the hard disk contains all of the sensitive data it contained before the erasure. So, the possible damage the data on the hard disk can cause can be approximated by the examples in chapter 9. Reading a small amount of hard disk sectors takes negligible time, less than one second in most cases. So, the cost of this kind of erasure verification is very low, but the possible benefit of detecting a failed erasure is immense.

Other kind of erasure errors are harder to detect, because only part of the hard disk is not erased. On the other hand, partial erasure means that less of the damaging data probably remains on the hard disk. In the chapter 11 there was a definition of type I large scale erasure errors. This kind of errors usually leave large parts of the hard disk unerased. The reason may be, for example, detecting the hard disk size incorrectly. For detecting this kind of erasure errors reading the last 100 sectors of the hard disk is somewhat effective. Most often type I errors cause only the end of the hard disk to not be erased, so it is likely that some data remains in the last 100 sectors. However, there are problems with this approach. When the hard disk is newly formatted, the operating system starts to fill it from the beginning of the hard disk. It is possible that the hard disk never gets so full that the last sectors are used. If this is the case, the last sectors are filled with zeroes because of the formatting, and reading them gives the false result of the hard disk being thoroughly erased even though some data remains on the hard disk.

For detecting type I large-scale errors random sampling from the whole hard disk is also needed. The smaller the amount of data remaining on the hard disk, the harder it is to detect it by sampling, but on the other hand, if only a little data remains it is likely that the damage it can do is smaller or that all damaging data is gone.

12.1 ERROR DETECTION BY SAMPLING

The mathematical model developed in chapter 11 assumes that erasure errors are independent of each other, and every sector has the same probability of containing the error. Because of the properties of arrival time process the errors probably cluster, but this does not mean that the probability of error is different in each part of the hard disk.

The reason why many applications of statistics rely on random sampling is that the samples differ from each other. For example in an election poll the random sample should be taken so that the sample is representative of the population. In case of a hard disk there is little difference between two sectors. It is not absolutely necessary that every sector has the same probability of being chosen to the sample. Spreading the samples evenly across the hard disk is much more important.

A solid choice would be reading every n th sector of the hard disk to make sure they do not contain data after the erasure. So, the verification scheme could be for example:

The 100 first sectors of the hard disk
The 100 last sectors of the hard disk
Every 200th sector in between the beginning and the end of the hard disk

Now the problem is how likely this erasure verification scheme finds the areas of failed erasure. To find out it is necessary to take a closer look at the model presented in chapter 11.

Large scale erasure errors tend to cover large areas of the hard disk. The model states that the unerased area begins at T_1. At this point a Poisson process with parameter λ_2 begins, so that when this arrival happens the unerased area ends. Large scale errors occur because of major problems in the erasure process; for example detecting the hard disk size wrong. For this reason the unerased area is probably quite large. If the hard disk size is 100 GB, which means that the hard disk contains 195358867 sectors, the λ_2 could be for example 0.000001, where λ_2 is measured in sectors.

Now it would be useful to find out how likely it is that the large scale erasure error process with parameter $\lambda_2 = 0.000001$ generates an area that is longer than 200 sectors. Çinlar notices (1975, 72) that for all $t \geq 0$ and constant $\lambda \geq 0$

$$P\{N_t = 0\} = e^{-\lambda t} \qquad \text{(Formula 12.1.1)}$$

For an arrival time process with λ = 0.000001 and t =200 holds true:

$$P\{N_t = 0\} = e^{-\lambda t} = e^{-0.0002} \approx 0.99980 \qquad \text{(Formula 12.1.2)}$$

So, we can say that a large scale error area is with very high likelihood longer than 200 sectors. Because the samples are taken every 200 sectors, an error area that is longer than 200 sectors is detected with 100% likelihood, provided the erasure verification software is "perfect". This is of course not the case in real world applications, but making the verification software as independent of the erasure software as possible gives the highest possible likelihood that the verification is able to detect the errors.

Detecting small scale erasure errors is much trickier, because in most cases they affect much shorter areas than 200 sectors. On the other hand, if a hard disk contains small scale erasure errors it most probably has lots of them, maybe thousands or tens of thousands. It is not necessary to find all of the erasure errors. Finding even one of them is enough for the person doing the hard disk sanitation to take special care of the hard disk, and destroy it by melting or pulverizing it.

According to the model developed in chapter 11.3, the length of a small scale erasure error is created by a waiting time process with λ about 1. From this it is easy to calculate that

$$P\{N_t = 0 \mid t = 1, \lambda = 1\} = e^{-1} \approx 0.3679 \qquad \text{(Formula 12.1.3)}$$

So, about 37% of the erasure failure areas are longer than one sector. Similarly it is easy to calculate

$$P\{N_t = 0 \mid t = 2, \lambda = 1\} = e^{-2} \approx 0.1353 \qquad \text{(Formula 12.1.4)}$$

$$P\{N_t = 0 \mid t = 3, \lambda = 1\} = e^{-3} \approx 0.0498 \qquad \text{(Formula 12.1.5)}$$

$$P\{N_t = 0 \mid t = 4, \lambda = 1\} = e^{-4} \approx 0.0183 \qquad \text{(Formula 12.1.6)}$$

Only 14% of the erasure failure areas are longer than two sectors, and less than 5% are longer than three sectors. If there are K such areas of failed erasure on the hard disk, how likely it is that at least one of these happens to be in one of the sectors which are verified?

A very similar problem is lifting balls from a basket that contains white and black balls. The white balls are sectors that are not verified, and the black balls are the sectors that are verified. The Poisson process that does the ball selection is random, even though the balls are neatly arranged inside the basket so that the black balls are an even distance from each other.

The distribution of white and black balls when K balls are lifted from the basket is hypergeometric. There are N balls total, of which M are black and $N-M$ white. If now X denotes the number of black balls among the K balls lifted, the probability can be calculated from the formula (Casella G, Berger L, 1990, 87)

$$P(X = x \mid N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}, \quad x = 0,1,\ldots,K \qquad \text{(Formula 12.1.7)}$$

The calculations are not easy when N and M are really large numbers. However, it is possible to use a normal approximation. Statistical Inference (Casella G, Berger L, 1990, 625) notes that for hypergeometric distribution

$$EX = \frac{KM}{N} \qquad \text{(Formula 12.1.8)}$$

$$VarX = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)} \qquad \text{(Formula 12.1.9)}$$

Now these can be used in the normal approximation with

$$\mu = \frac{KM}{N} \qquad \text{(Formula 12.1.10)}$$

$$\sigma^2 = \frac{KM}{N}\frac{(N-M)(N-K)}{N(N-1)} \qquad \text{(Formula 12.1.11)}$$

In case the hard disk contains 1000 small areas where the erasure has failed, $K = 1000$. If the hard disk is of 100 GB size, then $N = 195358867$ and $M = N/200 \approx 976794$.

Now the approximation has the following parameters

$$\mu = \frac{KM}{N} = \frac{1000 * 976794}{195358867} \approx 5 \qquad \text{(Formula 12.1.12)}$$

$$\sigma^2 = \frac{KM}{N} \frac{(N-M)(N-K)}{N(N-1)} = 5 * \frac{194382073 * 195357867}{195358867 * 195358866} \approx 4.97497 \qquad \text{(Formula 12.1.13)}$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{4.97497} \approx 2.2304$ (Formula 12.1.14)

On average the sampling would find 5 small areas of failed erasure, and the likelihood that at least one was found is

$$P\left(X \geq \mu - \frac{4}{2.2304} \sigma \right) = P(Z \geq -1.7934) \approx 0.9633, \quad Z \sim n(0,1) \qquad \text{(Formula 12.1.15)}$$

So, if the hard disk contained 1000 areas where the erasure has failed and each would be exactly one sector long the verification scheme would detect the error with over 95% probability. In fact the likelihood that the error is detected is even higher, because some of the areas are bigger than one sector.

If the amount of sectors with failed erasure is much lower the detection strength of the verification scheme outlined in chapter 12.1 is lower. However, this erasure verification model is very good in making sure that most of the hard disk has been successfully erased. A perfect verification software following the verification scheme would detect a catastrophic erasure error practically every time. Large erasure errors would be detected almost always. Small erasure errors would be detected if there were enough of them. Generally, if the erasure verification scheme of verifying the 100 first and last sectors and every 200th sector in between does not find any data the hard disk is probably fully sanitized or only a few areas of failed erasure remain.

12.2 LIKELIHOOD THAT ANY SENSITIVE DATA REMAINS AFTER ERASURE

Chapter 11 lists different kind of erasure failures: large scale erasure errors, small scale erasure errors and catastrophic erasure errors. If sanitating the hard disk fails it is likely that some data remains on the hard disk. In case of catastrophic erasure error all of the original data remains. However, if the erasure error is detected in the verification process the hard disk can be disposed of safely. An interesting problem is how much data can remain on the hard disk after the erasure, if the erasure verification states that the erasure was successful, and how likely it is that this amount of data contains some sensitive material.

If the erasure verification scheme used is “first and last 100 sectors and every 200th sector in between”, any large unerased area should be found by the verification process. This holds true if the verification is perfect, i.e. does not suffer from the same limitations as the erasure software. In practice this is not always the case, because most often erasure and verification are done by different parts of the same software.

If the erasure verification works as intended and reports that the hard disk is fully sanitized, in most cases the hard disk is fully erased. Some small areas can escape the verification, but how likely it is that these contain sensitive data? The undetected areas of failed erasure are shorter than 200 sectors, because any longer area would be detected by the verification. The number of the areas is low enough that they escape the verification process.

The small scale erasure error model can be used to simulate a case where the errors do not get detected. If the hard disk size is 100 GB, which means it contains 195358867 sectors, a possible arrival time process that generates the erasure errors could be a Poisson process with λ = 0.000005. When each of these arrivals happens, another process starts with λ_2 = 1. Now the areas of failed erasure lie in between these arrival points.

$$E[T_n] = \frac{n}{\lambda} = \frac{195358867}{0.000005} \approx 1000 \qquad \text{(Formula 12.2.1)}$$

The model would generate on average 1000 areas of failed erasure, most of which are one sector long. This can happen with a hard disk that is slightly broken. As formula 12.1.15 shows the erasure verification finds the errors with over 95% likelihood, so in case the hard disk contains erasure failures that are not detected it is likely that the number of unerased sectors is under 1000.

If the hard disk contains 1000 unerased sectors out of 195358867, then about one sector in 200000 remains intact after the erasure. These remaining sectors are most likely spread around the hard disk, because large unerased areas would be easily detected. Most likely the hard disk was not full of sensitive data before the erasure. If for example 100 MB of the data on the hard disk is really sensitive, then on average one sector out of 1000 contains sensitive data. The amount of unerased sectors with sensitive data can be calculated with normal approximation of hypergeometric distribution, with parameters N = 195358867, M = 195359 and K = 1000.

$$\mu = \frac{KM}{N} = \frac{1000 * 195359}{195358867} \approx 1 \qquad \text{(Formula 12.2.2)}$$

$$\sigma^2 = \frac{KM}{N} \frac{(N-M)(N-K)}{N(N-1)} = 1 * \frac{195163508 * 195357867}{195358867 * 195358866} \approx 1 \qquad \text{(Formula 12.2.3)}$$

Standard deviation $\sigma = \sqrt{\sigma^2} = \sqrt{1} = 1$ (Formula 12.2.4)

Now the amount of sectors with sensitive content is with 99% likelihood lower than $\mu + 2.33\sigma$, so most likely the hard disk contains 4 or less sectors with sensitive data. This can happen for example if one of the unerased areas is 4 sectors long and coincides with sensitive content.

The 4 sectors contain 2 KB of data, which can hold a very short plain text file of about 200 words. Any more complex files do not fit in such a small space. If the possibility that this kind of short pieces of text can remain is a major problem, then choosing a high-security erasure method with full disk verification is recommended. If the data on the hard disk is not so sensitive the proposed erasure verification is enough.

13 OPTIMAL ERASURE VERIFICATION SCHEME

The optimal erasure verification depends on the sensitivity of the data on the hard disk. In the case the data is top secret or very damaging, it is prudent to verify the whole hard disk. This takes about the same time as a single overwriting round. Most banks and other high-security installations overwrite 3 or 7 times, so the time cost of full verification is not too high compared to the time required for erasure.

If the data on the hard disk is not as sensitive, the time cost of erasure verification becomes more important. The erasure verification should take only a short time but give a high certainty that the data is erased. A good choice for erasure verification scheme would be verifying:

a) The 100 first sectors of the hard disk
b) The 100 last sectors of the hard disk
c) Every 200th sector in between the beginning and the end of the hard disk

This gives a high certainty that the erasure was successful. Even if some sectors fail to be erased, the amount of unerased sectors is probably so low that only negligible amounts of sensitive data remain.

The time required for this kind of erasure verification is approximately $^1\!/_{200}$ of the time required for one erasure round. When an average single erasure round with most hard disks takes about one hour, the erasure verification would take about 20 seconds. This kind of time cost is acceptable.

14 DISCOURSE ABOUT THE RESULTS

This thesis is based on the idea of imperfect hard disk sanitizing software and perfect erasure verification software. In practice both of them are imperfect. Some of the erasure errors are easy to detect even with imperfect verification software; for example a catastrophic erasure error that leaves the whole hard disk unerased can usually be detected just by reading the first 100 sectors at the beginning of the hard disk. Most often these sectors contain data if the hard disk has been in use. However, sometimes the problem that caused the erasure error also makes the erasure verification fail.

One example is detecting the hard disk size wrong because the BIOS reports the hard disk as smaller than it really is. If the same programmer has designed both the erasure software and the verification software, it is likely that they both use the same kind of program code to detect the hard disk size. If both erasure software and verification software detect the hard disk size wrong it is likely that the verification does not detect that a large part of the hard disk is left unerased.

Another problem that makes trusting the verification software hard is that the hard disk itself has a controlling circuit that can for example detect broken sectors and relocate the data to remapped sectors, in effect leaving the data that is in the broken sector intact. In this case the software would report that all of the sectors are erased properly, even though some data may remain in the remapped sectors. It is hard to develop verification that gets around this problem.

In practice erasure verification detects most erasure errors, but even verifying the full hard disk does not give 100% certainty that no data remains unerased. The problem will get even worse when hardware develops further, and software commands for manipulating the data on the hard disk have to work through command interpretation layers. Direct ATA commands may not do what the programmer thinks they should do.

Before this thesis no scientific papers existed on verifying hard disk erasure. As this is the first thesis of its kind it only lays the basics on erasure verification. Further studies are required to fully understand the problem.

CITATIONS

Aquisti, A, Friedman, A. & Telang, R. 2006. Is There a Cost to Privacy Breaches? An Event Study. Twenty-Seventh International Conference on Information Systems, Milwaukee

BBC News, Nationwide laptop theft 2006. Security raised over laptop theft (18.11.2006) <http://news.bbc.co.uk/2/hi/uk_news/6160800.stm> (read 29.4.2007)

Boran, O, 2004. Basic Data Collection on E-Waste Recycling in Yemen. Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) GmbH, Sana'a

Cavusoglu, H., Mishra, B & Raghunathan S. 2004. The Effect of Internet Security Breach Announcement on Market Value: Capital Market Reactions for Breached Firms and Internet Security Developers. International Journal of Electronic Commerce, Fall 2004, Vol. 9, No. 1

Casella, G, Berger, R 1990. Statistical Inference. Brooks/Cole publishing company, Belmont, California

Çinlar, E. 1975. Introduction to Stochastic Processes. Northwestern University, New Jersey

Computer Industry Almanac Inc 2006. 25-year PC Anniversary Statistics. <http://www.c-i-a.com/pr0806.htm> (read 30.3.2008)

CMS Information Security Levels 2002. National Institute of Standards and Technology (NIST) Computer Security Resource Center (CSRC) <http://csrc.nist.gov> <http://csrc.nist.gov/fasp/FASPDocs/risk-mgmt/ssl.pdf> (read 17.3.2007).

Dew Associates Corporation Knowledge Center: Hard Disk Sector Structures <http://www.dewassoc.com/kbase/hard_drives/hard_disk_sector_structures.htm> (read 30.5.2008)

DoD 5220.22 2004. United States of America Department of Defense Directive 5220.22, Washington, DC

DSS Cleaning & Sanitation Matrix 2005. The Defense Security Service (DSS) Office of Designated Approving Authority (ODAA) <http://www.dss.mil/infoas/> <http://www.dss.mil/files/pdf/clearing_and_sanitization_matrix.pdf > (read 17.3.2007).

Geoff, G, 2007. The Tech Report: Seagate's Barracuda 7200.11 hard drive. <http://techreport.com/articles.x/13440/1> (read 30.3.2008)

Grant, A, Meadows, J 2006. Communication Technology Update, 10th Edition. Elsevier Science & Technology, Washington, DC

Grochowski, E., Halem, R. 2003. Technological impact of magnetic hard disk drives on storage systems. IBM Systems Journal, Volume 42, Number 2, 2003. <http://www.research.ibm.com/journal/sj/422/grochowski.html> (read 1.4.2008)

Gupta, M., Hoeschele, M. & Rogers, M. 2006. Hidden Disk Areas: HPA and DCO. International Journal of Digital Evidence. Fall 2006, Volume 5, Issue 1. <http://www.utica.edu/academic/institutes/ecii/publications/articles/EFE36584-D13F-2962-67BEB146864A2671.pdf> (read 30.3.2008)

Gutmann, P, 1996. Secure Deletion of Data from Magnetic and Solid-State Memory. Proceedings of Sixth USENIX Security Symposium, San Jose, California. <https://www.usenix.org/publications/library/proceedings/sec96/full_papers/gutmann/> (read 7.5.2007)

HMG Infosec Number 5, 2003. HMG Infosec Standards 5, London.

Matthews, H., McMichael, F., Hendricson, C. & Hart, D. 1997. Disposition and End-of-Life Options for Personal Computers. Green Design Initiative Tech Report #97-10. Carnagie Mellon University, Pittsburgh, PA.

Microsoft Windows NT Server Resource Guide, chapter 3 – Disk Management Basics. <http://www.microsoft.com/technet/archive/winntas/support/diskover.mspx?mfr=true> (read 30.5.2008)

Nicholls, S, Kushin, M 2006. FRP-Reuseit – Recycling and Re-using ICT Equipment. <http://www.icthubknowledgebase.org.uk/recyclingict> (read 30.3.2008)

Storagereview.com Reference Guide – Hard Disk Drives 2005. Hard Disk Size Barriers. <http://www.storagereview.com/guide2000/ref/hdd/bios/size.html> (read 30.3.2008)

University of Utah Metallurgical Engineering Gallery: Computer Hard Disk Picture <http://www.metallurgy.utah.edu/galleries/hardisk3.jpg> (read 30.3.2008)

U.S. Census Bureau 2008. The 2008 Statistical Abstract. Table 624: Mean Hourly Earnings and Weekly Hours by Selected Characteristics: 2005, Washington, DC. <http://www.census.gov/compendia/statab/tables/08s0624.pdf> (read 24.2.2008)

Valli, C. 2004. Throwing out the Enterprise with the Hard Disk. Edith Cowan University, Western Australia.