**Abstract**

**1. Introduction**

Currently, there is a huge push for “big data” and for the analysis of “big data”. This is resulting in massive amounts of data being stored and used. Inevitably, a certain percentage of data stored will be corrupted due to disk failure and other factors. While this percentage is effected by choice of file system, the fact remains that data is reliably corrupted at a rate that is above zero. As more data is stored and accessed for data mining and other “big data” research, the odds of needing data that has been corrupted increases. For this reason, we saw a need to study the effects that hardware corruption, emulated by random bit flipping, has on different file types in order to determine if certain file types are more resistant to corruption than others. In this paper, we use random bit flipping to emulate the corruption caused by hardware failures. We seek to examine the fault tolerance of different popular file types in order to study the resistance of various file types to hardware failures that result in bit flipping errors.

In order to accomplish this task, we developed a program that would read in a file and flip a specified number of bits ranging from 1 to 1,000. The specific bits that were flipped were randomly selected using a random number generator. After processing a file through our “bit flipper”, we ranked it by usability. We performed many repetitions of this and various file types, and so were able to draw conclusions on which file types and sizes were more resistant to hardware failure resulting in file corruption. Our full methodology is documented in section 4.

**Contributions:** We make the following contributions:

We created a program to emulate the effects of hardware failure on a variety of file types

We establish a rating scale to score the usability of a corrupted file

We detail the resistance of different file types and sizes to corruption due to hardware failure

**2. Background**

**3. Prior Studies**

Prior studies on corruption have focused not on the effects of it, but on detecting and repairing corruption.

<brief overview of studies detecting and repairing corruption>

This paper does not focus on detecting or repairing corruption, but simply on what corruption looks like in various file types and how usable they are after they are corrupted.

**4. Methodology**

We chose to examine the following file types:

.txt

.doc

.docx

.pdf

.png

.bmp

.jpg

.exe

For each file type, we tested in multiple files ranging in length and/or complexity. For each individual file, we flipped 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 50, 100, 500, and 1000 random bits. For each number of bits flipped, we produced ten corrupted files. Correspondingly, each file that went into the bit flipper resulted in the output of 170 different corrupted files: 10 files with 1 bit flipped, 10 files with 2 bits flipped, 10 files with 3 bits flipped, and so on.

In this paper, a test case refers to the combination of a file and a number of bits flipped. We chose to run each test case ten times. Clearly a single file cannot be representative of an entire test case. Similarly, the more files we examine for each test case the more accurate our usability score will become. We chose to run each test case ten time so that we could get an accurate representation of what the average test case would look like, while still keeping the scale small enough that we would be able to visually inspect each file.

Each file that the bit flipper output was visually examined and scored from 0 to 5, with 0 being unopenable and 5 being perfect. Further details on scoring can be found in section 4.2. For each test case, an average usability score was calculated and this average score is the value referred to as the **usability score** for the duration of this paper. As the usability score is a mean and a mean is not always the best measure of central tendency, there are instances where we will delve further into the data and judge it in different ways.

In addition to calculating the usability score for each test case, we kept track of the percent of files that opened for each test case. As we ran each test case ten times, this number is always a nice round percentage.

**4.1 Flipping Bits**

<code snippet and discussion>

<adam>

**4.2 Scoring Files**

Every file was ranked from 0 to 5 as follows:

0 – The file does not open at all. A window may open, but no part of the file loads.

1- The file opens but is corrupted beyond usability.

2- The file opens and is obviously corrupted, but recovery of some data is possible.

3- The file opens and nearly all data is intact, but some corruption is easily detected.

4- The file opens and most data is intact, a diff is required to find any change in the file.

5- The file opens and even a diff cannot detect a change.

While the above scoring system guided all efforts, further descriptions will be given for each file type examined and how the test files of the specific file type were judged to fall into a given score.

There is more granularity in this score system for files that function well than for those that function poorly. We briefly reviewed the test files before we developed the rating scale and we noticed a need for more granularity among files with a high level of usability.

An important aspect of this scoring system is that we focused not on some percentage of characters changed or a purely quantitative measure, but on the actual usability of the document. This is something of a rarity in computer science, but we felt it was more important to understand how useful the corrupted document still is than to generate statistics on how much of the document changes.

As there were several thousand files to be examined, we divided the scoring among the three members of the research group. In order to ensure consistency of scoring within a single file type, a single person scored all files for a given file type. All document formats were handled by a single person, all executables by a different person, and the image files were split among two people, but a single person handled all of a given file extension.

**4.3 Methods Limitations**

We did not examine cross-scorer validity <This isn’t the name for it. Having a brain fart… >

The subjective nature of our scoring system leaves blah blah blah

There is not a fine level of granularity in our scoring system. Two different test cases scored as two could still look vastly different as it covers all files that function and allow for the recovery of some data through files that are clearly corrupted.

**5. Results**

**5.1 .txt Files**

.txt files were unique in that all corrupted files could still be opened. This is likely the result of a small header that that rarely got hit by a random bit flip, and which didn’t affect the file’s ability to be opened if it did hit the header. They were also unique in that every single file showed some evidence of corruption. While the evidence was frequently very small and could only be detected by using a diff, every file had some change in the viewable text from the original file.

Text files were also unique in that they were the only file type where we saw a complete change in character set. This occurred several times in the small files. When this occurred, the file was scored as a 1.

It was readily apparent in text files that the smaller the file was, the more rapid the deterioration of usability. This observation is logical, as given that the number of flipped bits holds consistent, the smaller the file size, the higher the density of the flipped bits. As the small file was only five words, it deteriorated very quickly. The medium file was five pages, which is quite a jump from 5 words, and while it did deteriorate slightly more quickly than the large file, it was still a very slow deterioration and not nearly as dramatic as the deterioration of the small file.

When viewing the text files, all files fell into the range of 2-4, with the vast majority of files being scored as threes or fours. The small files deteriorated down to a usability score of one, but the medium and large files both maintained very high functionality, never even falling below a usability rating of three. The medium and large files were still completely readable, with only misspellings and non-alphanumeric numbers serving as evidence of their corruption. Up until about 50 bits flipped, a diff was frequently required to find any difference between the corrupted and non-corrupted files in the small and large test files.

**5.2 .doc files**

The performance of the .doc files surprised the research group. We expected the .docs to fail quickly, but were surprised by the number of the corrupted files that could be opened. As opposed to the text files where all files could be opened and none of them were perfect, not all of the .doc files could be opened and there were multiple examples of perfect files. In fact, many of the small .doc files with only a few bits flipped received a usability score of 5, meaning that we could not detect, even with a diff, any change in the document. This would indicate that the bit flip occurred in some part of the header that was not accessed.

Looking only at the average usability scores, all three file sizes rank quite closely to each other in terms of usability score. The small file starts with a higher usability, but by the point three bits are flipped, the files fall into the expected order – with the large files being the most usable and the small files the least. The files mostly stay in that order as more and more bits are flipped, with the ordering becoming more consistent as more bits are flipped. Looking only at the average scores fails to show some interesting trends in the performance of the .doc files.

The small .doc test files tended to score either a score of five or of zero, with only a few files falling in between. The small .doc files were also the only .doc files where test files scored fives. Additionally, it wasn’t just one or two test files that scored fives, but all ten files with one bit flipped, nine with two bits flipped, and six with three bits flipped. This is quite a difference from the medium and large files, where not a single file was perfect. I suspect that this is because due to the lack of content in the small files, the bit blip was most likely to fall within the header where it would either make the file completely unreadable, or wouldn’t affect the file at all. Figure <blah> shows the rate at which files failed to open for the small, medium, and large .doc files. You can see that the small files fell of more quickly than the medium and large files did. In conclusion, small files were more likely to function perfectly and more likely to fail to open than both medium and large files.

While the medium and large .doc files did not have any examples of perfect files, there were plenty of examples of high functioning files. If the file opened, it was almost always scored either a 3 or a 4, meaning it was easily readable and any missing or changed data could be inferred. As mentioned when discussing the functioning of the small files, the medium and large files were both more likely to open and significantly less likely to function perfectly than small files.

**5.3 .docx files**

The .docx files were the easiest to score, as very few of them opened. Any docx file that did open was completely perfect and scored a 5.

Among the small files, only two files opened, both of them having only one bit flipped. This resulted in an average usability score of 1 for the small.docx files with 1 bit flipped. This is again an instance in which the mean is a bad measure of central tendency as it overestimates 80% of the data and strongly underestimates 20% of the data. Perhaps more revealing is that the median and the mode are both zero and the standard deviation is 2. The usability score for all tests beyond one bit flipped is zero.

For both the medium and large files, only a single file opened. In both instances, the single file was one of the ten with only bit flipped. This resulted in a usability score of .5 for a single bit flipped, and of 0 for all other cases.

**5.4 .pdf files**

PDF files were the most interesting to score as they were the only file type where flat out strange things happened. Fonts changed from page to page, spacing and formatting were off, and entire blocks of text would disappear.

Anytime a PDF appeared to be perfect and I ran it through a diff, it turned out to be perfect. When a PDF corrupted, it wasn’t subtle about it. Either formatting obviously changed or a block of text disappeared, leaving a big blank space. This means that there wasn’t a single PDF with a score of 4.

Another factor that was unique to PDFs was the number of errors that they threw. Scoring the PDFs took significantly more time, simply because for a single document, I may have had to click through 15-20 errors in order to see the whole file. It seemed that it notified me that it detected corruption not once for the entire file, as Microsoft Word did, but each time that it detected some corruption as I scrolled through the file.

PDFs were the only file type that would open a window and present a blank document, but fail to present any text from that document. When this occurred, the file was marked as if it did not open at all as there was no text of any sort rendered.

By usability score, PDF files deteriorated more consistently than others and the small, medium, and large files deteriorated at roughly the same rate. Figure <BLAH> shows the rate of deterioration of PDF files with 1-10 bits flipped. This finding is in contrast to the other file types where small deteriorated significantly more quickly than medium and medium deteriorated more quickly than large. While the usability scores fail to show small deteriorating significantly more quickly, the open rates do. As show in <FIGURE BLAH>, even from a single bit flipped, the small PDFs are less likely to open than the medium or large ones.

**5.5 …**

**5.? Comparison of File Types**

<graph plotting all – usability scores>

<graph plotting all – open rate>

**6. Implications**

**6.1 Future Research / Unanswered Questions**

**7. Conclusions**

**References**