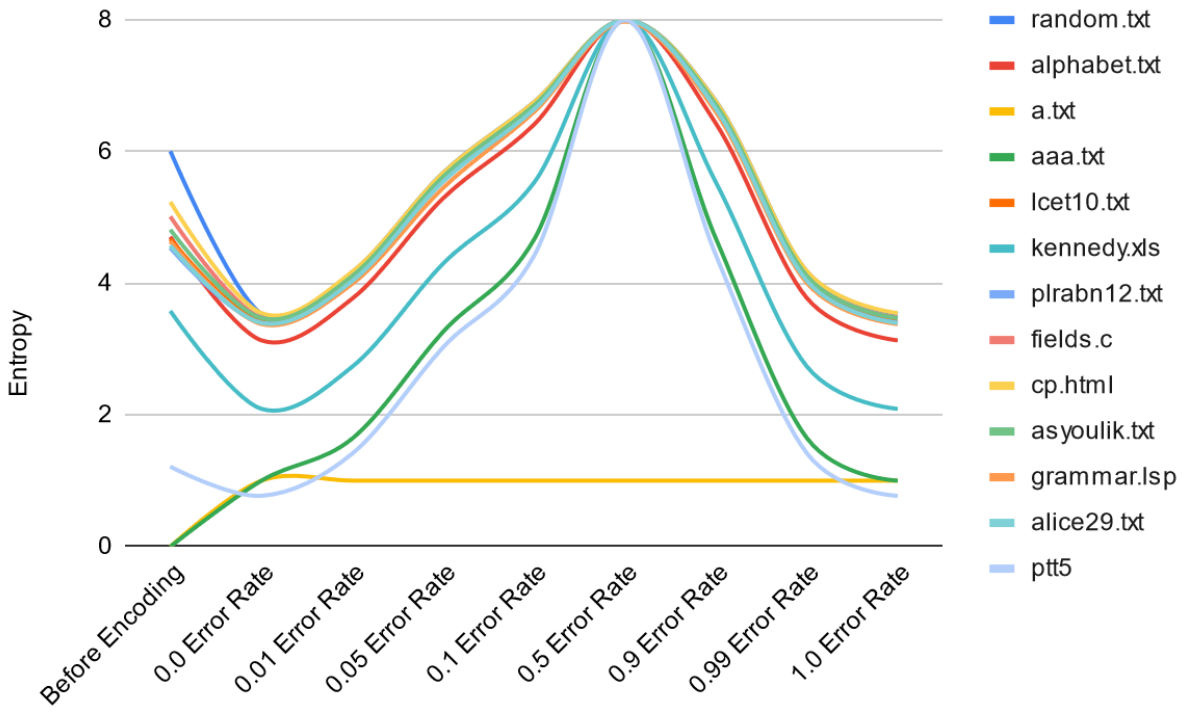
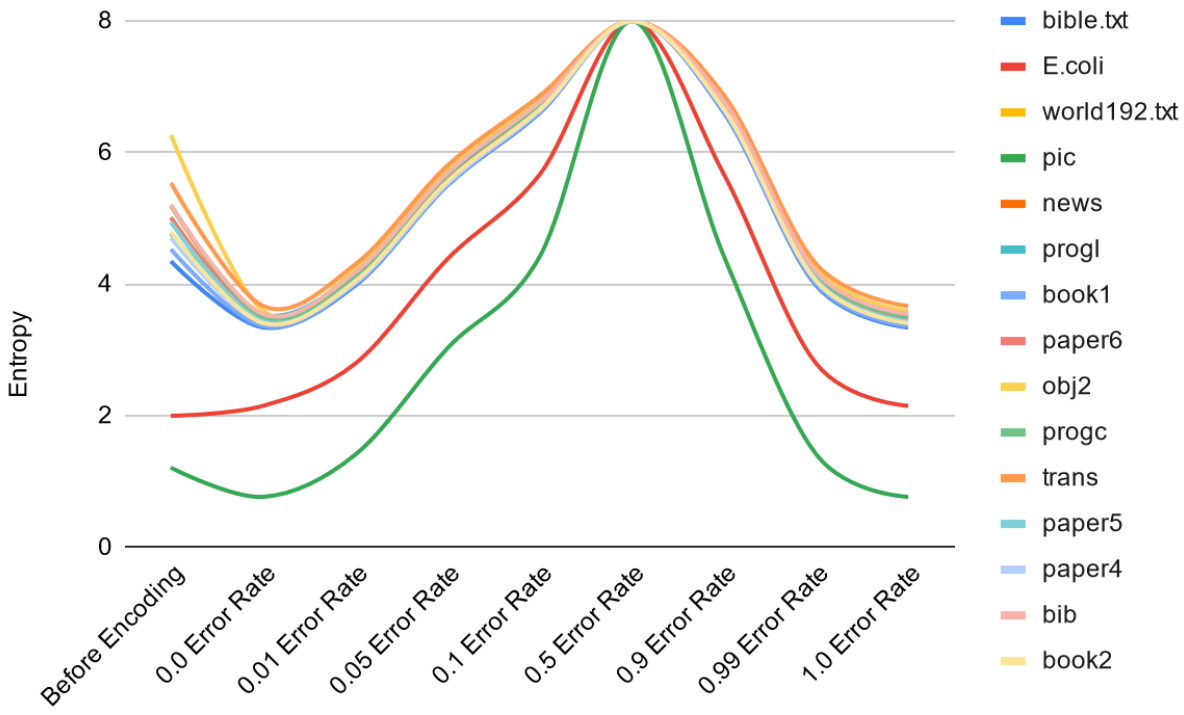
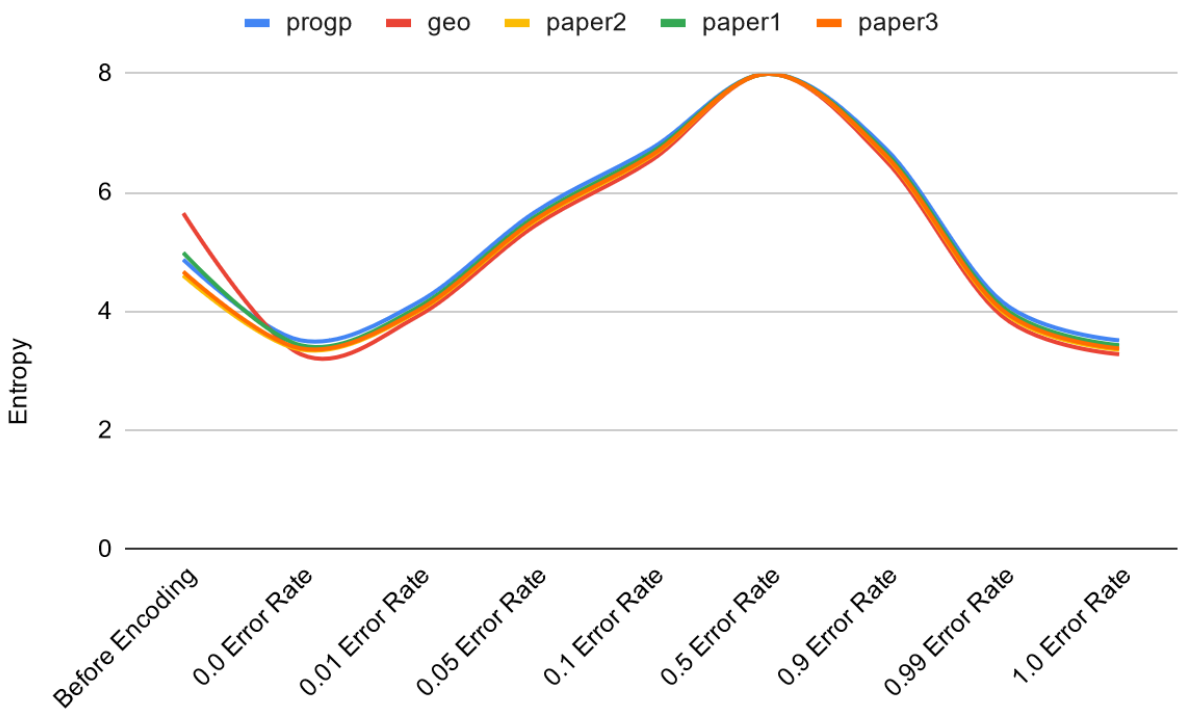
Darren Lee
CSE13S

Assignment 5: Hamming Codes
Writeup



Graph of entropy before encoding and after encoding with varied error rates


8
6
4
2
0
Entropy
Before Encoding
0.0 Error Rate
0.01 Error Rate
0.05 Error Rate
0.1 Error Rate
0.5 Error Rate
0.9 Error Rate
0.99 Error Rate
1.0 Error Rate
bible.txt
E.coli
world192.txt
pic
news
progl
book1
paper6
obj2
progc
trans
paper5
paper4
bib
book2


progp
geo
paper2
paper1
paper3
8
6
4
2
0
Entropy
Before Encoding
0.0 Error Rate
0.01 Error Rate
0.05 Error Rate
0.1 Error Rate
0.5 Error Rate
0.9 Error Rate
0.99 Error Rate
1.0 Error Rate

Analysis

Entropy is a measure of randomness and uncertainty. For the files in these graphs, the entropy after encoding with a 0.0 error rate is often less than the entropy before encoding. This is because encoded Hamming Codes have some predictability and repetitiveness. With only 4 bits to represent the code, there are only 2^4=16 possible parity bit codes with a Hamming(8,4) code. Therefore, there will be quite a few repetitive portions of an encoded message. Unless the data was already very repetitive in the first place, such as with a.txt, aaa.txt, or E.coli, encoding with a 0.0 error rate will generally decrease the level of entropy in the data. This is because there is very little uncertainty when data is repetitive, reducing the level of randomness in the data. In a file such as a.txt, where there is only the letter a, there is very little uncertainty even after encoding with small and large error rates. This is because there is only one character, represented by 1 byte, in the a.txt file. There is only one byte to encode, and there is not much uncertainty even after encoding with many errors because there are not that many bits to flip in the first place. Entropy tends to peak at an error rate of 0.5 and then steadily declines as error rates grow larger than that. Using $\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log \mathrm{P}(x_i)$, where x_i represents possible outcomes and $P(x_i)$ represents their probabilities, we can see that entropy is lowest when there is only one event with absolute certainty and grows largest when events have equal probability of occurring. Knowing this, it makes sense that entropy is small when many bits are not flipped or many bits are flipped because there is a larger probability of an event happening or not happening. With an error rate of 0.5, there is theoretically an equal likelihood of flipping a bit or not flipping a bit, so entropy will be highest with such an error rate. The message will be considerably different from what it originally was, but the probability of bit flipping is more likely with a high error rate. This means there is a greater degree of certainty in the data. This is

reflected in the general trend of the data to spike in entropy when the error rate is half way from 0 to 1 but dip lowest near the ends at 0 and 1.