DAT470 Assignment 4

Olof Lindberg
Rikard Roos

March 2025

Problem 1

b)

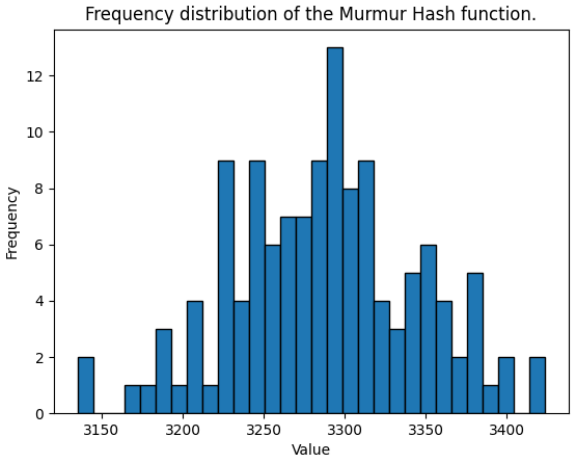


Figure 1: Frequency distribution of the distribution of hash values.

Total number of words: 420769

Mean: 3287

Standard deviation: 57

Collision probability = $P[h(x) = h(y)|x \neq y] = 0.007813 \approx 0.78\%$ (estimated by $\frac{\#\text{collisions}}{\#\text{key pairs}}$)

c)

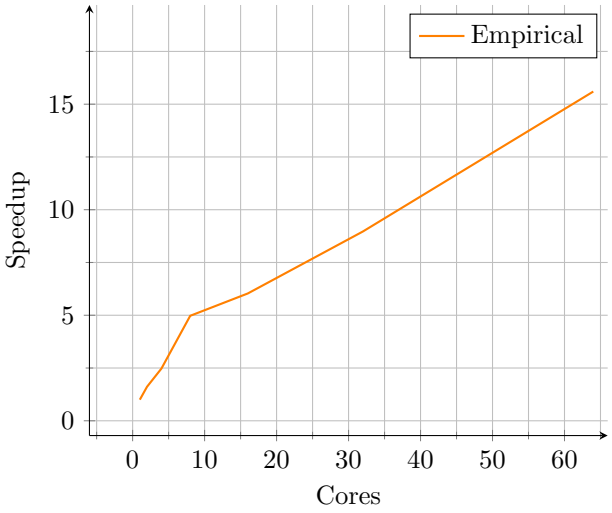
The hash function seems to be performing good in the sense that it somewhat uniformly distributes the data. With an average of 3287, the standard deviation of 57 could be considered to be very low, indicating a uniform frequency distribution. Furthermore, a perfectly uniformly distributed hash function with 128 buckets should have a collision probability of $\frac{1}{128} = 0.78125\%$. The probability estimation of 0.7813 % for our Murmur Hash function further indicates a good hash function.

Problem 3

a)

Estimate for the huge dataset with m = 1024 registers and the seed 0x9747b28c: 52099984

b)



c)

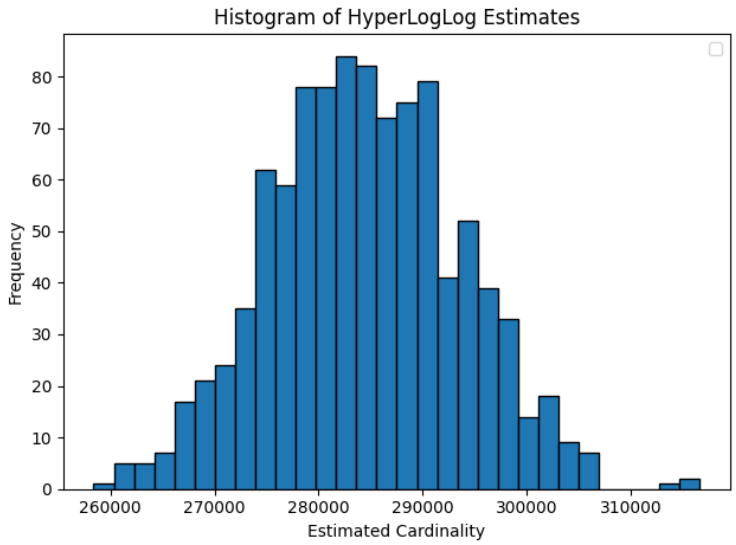


Figure 2: histogram of estimates

Metric	**Value**
True count (n)	284689
Average estimate	284426.52
Standard deviation	9077.17
Fraction within $n(1 \pm 1\sigma)$	68.80%
Fraction within $n(1 \pm 2\sigma)$	96.40%
Fraction within $n(1 \pm 3\sigma)$	99.70%

Table 1: Summary of HyperLogLog Estimations with $m = 1024$ and 1000 random seeds.