# Simulation Report

ziye.tao

June 2020

## 1 Result

First recall that $A$ represents the number of all patients, $B$ represents the number of patients who meet some criteria and $m$ is the number of buckets in the HLL process. We introduce $r := \frac{B}{A}$ to represent the ratio of $A$ and $B$. The purpose of running simulations under different combinations of $A$,$r$ and $m$ is to construct a table to fit Approximation1 and Approximation2 under these combinations. In all simulations, I restrict $A$ in the interval $[10^4, 10^7]$ and $m$ in the interval $[100, 50000]$. Since the simulations are run under the condition of 10-anonymity, I make sure that $\frac{A}{m} > 20$ which is the mean value of the single bucket size. Also, $r$ is restricted in the interval $[0.005, 0.1]$ and I choose 5 different $r$ which are: $0.1, 0.08, 0.05, 0.01, 0.005$ to run the simulations and compare the simulation results with computing results.

The final choice of Approximation1 and Approximation2 is mainly based on $\frac{A}{m}$. In most cases, when $\frac{A}{m} \geq 1000$, Approximation2 is good enough and the computing time will no longer than 3 minutes. When $\frac{A}{m} < 1000$, Approximation2 will be not accurate and we have to choose Approximation1. The computing time of Approximation1 which is proportional to $\sqrt{r}\frac{A}{m}$ is the main concern, and when $\frac{A}{m} < 1000$, the computing time is usually no longer than 15 minutes. But there is several special cases, when $r = 0.005$, the threshold of $\frac{A}{m}$ will be increased to 1500. But fortunately, the computing time of both Approximation1 and Approximation2 are proportional to $\sqrt{r}$ which makes sure that the computing time of Approximation1 no longer than 5 minutes when $\frac{A}{m} < 1500$ and $r = 0.005$.

# 2 Table

| | | | r = 0.1 | | r=0.08 | | r=0.05 | | r=0.01 | | r=0.005 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\frac{A}{m}$ | A | m | A1 | A2 | A1 | A2 | A1 | A2 | A1 | A2 | A1 | A2 |
| 100 | $10^4$ | 100 | √ | | √ | | √ | | √ | | √ | |
| 50 | $10^4$ | 200 | √ | | √ | | √ | | √ | | √ | |
| 20 | $10^4$ | 500 | √ | | √ | | √ | | √ | | √ | |
| 1000 | $10^5$ | 100 | | √ | | √ | | √ | | √ | | √ |
| 500 | $10^5$ | 200 | √ | | √ | | √ | | √ | | √ | |
| 200 | $10^5$ | 500 | √ | | √ | | √ | | √ | | √ | |
| 100 | $10^5$ | 1000 | √ | | √ | | √ | | √ | | √ | |
| 50 | $10^5$ | 2000 | √ | | √ | | √ | | √ | | √ | |
| 20 | $10^5$ | 5000 | √ | | √ | | √ | | √ | | √ | |
| 10000 | $10^6$ | 100 | | √ | | √ | | √ | | √ | | √ |
| 2000 | $10^6$ | 500 | | √ | | √ | | √ | | √ | | √ |
| 1500 | $10^6$ | 666 | | √ | | √ | | √ | | √ | | √ |
| 1000 | $10^6$ | 1000 | | √ | | √ | | √ | | √ | √ | |
| 500 | $10^6$ | 2000 | √ | | √ | | √ | | √ | | √ | |
| 200 | $10^6$ | 5000 | √ | | √ | | √ | | √ | | √ | |
| 100 | $10^6$ | 10000 | √ | | √ | | √ | | √ | | √ | |
| 50 | $10^6$ | 20000 | √ | | √ | | √ | | √ | | √ | |
| 20 | $10^6$ | 50000 | √ | | √ | | √ | | √ | | √ | |
| 100000 | $10^7$ | 100 | | √ | | √ | | √ | | √ | | √ |
| 20000 | $10^7$ | 500 | | √ | | √ | | √ | | √ | | √ |
| 10000 | $10^7$ | 1000 | | √ | | √ | | √ | | √ | | √ |
| 5000 | $10^7$ | 2000 | | √ | | √ | | √ | | √ | | √ |
| 3333 | $10^7$ | 3000 | | √ | | √ | | √ | | √ | | √ |
| 2000 | $10^7$ | 5000 | | √ | | √ | | √ | | √ | | √ |
| 1500 | $10^7$ | 6666 | | √ | | √ | | √ | | √ | | √ |
| 1000 | $10^7$ | 10000 | √ | | √ | | √ | | √ | | √ | |
| 500 | $10^7$ | 20000 | √ | | √ | | √ | | √ | | √ | |
| 200 | $10^7$ | 50000 | √ | | √ | | √ | | √ | | √ | |

Note: A1 and A2 represents Approximation1 and Approximation2.

For all choices of r,the longest computing time occurs at $A = 10^7$.