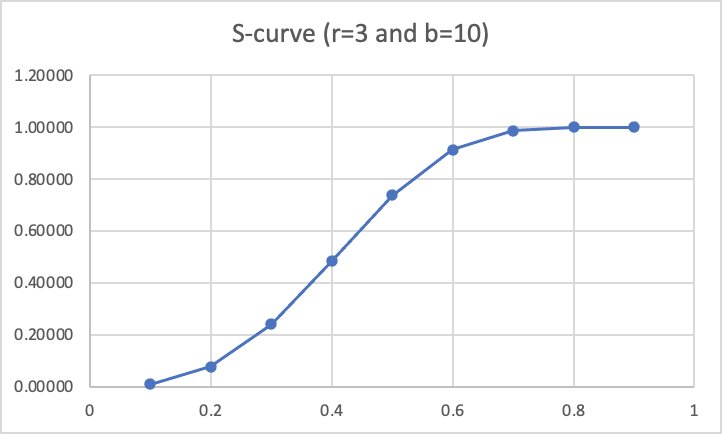
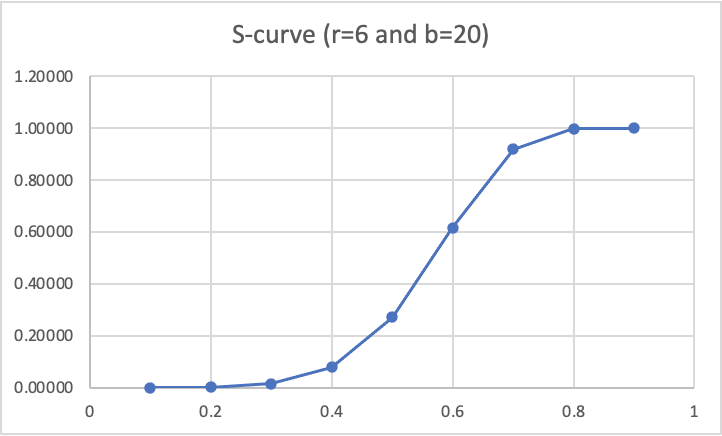
### MBD Assignment 2 - Team 6

Yingyao Lu a1784870 Jin Yin aXXXX

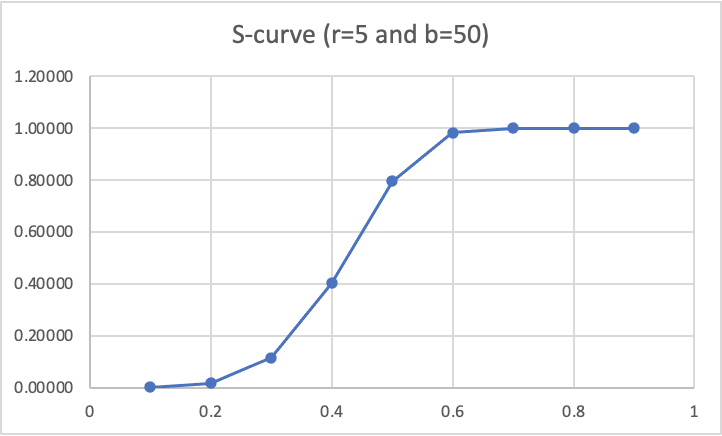
## Exercise 1 S-curve (5+5+5 points)

1.S-curve for r=3 and b=10

|  |  |
| --- | --- |
| s | S-curve (r=3 and b=10) |
| 0.1 | 0.00996 |
| 0.2 | 0.07718 |
| 0.3 | 0.23945 |
| 0.4 | 0.48387 |
| 0.5 | 0.73692 |
| 0.6 | 0.91227 |
| 0.7 | 0.98502 |
| 0.8 | 0.99923 |
| 0.9 | 1.00000 |

2. S-curve for r=6 and b=20

|  |  |
| --- | --- |
| s | S-curve (r=6 and b=20) |
| 0.1 | 0.00002 |
| 0.2 | 0.00128 |
| 0.3 | 0.01448 |
| 0.4 | 0.07881 |
| 0.5 | 0.27019 |
| 0.6 | 0.61541 |
| 0.7 | 0.91819 |
| 0.8 | 0.99771 |
| 0.9 | 1.00000 |

3. S-curve for r=5 and b=50

|  |  |
| --- | --- |
| s | S-curve (r=5 and b=50) |
| 0.1 | 0.00050 |
| 0.2 | 0.01588 |
| 0.3 | 0.11454 |
| 0.4 | 0.40228 |
| 0.5 | 0.79555 |
| 0.6 | 0.98253 |
| 0.7 | 0.99990 |
| 0.8 | 1.00000 |
| 0.9 | 1.00000 |

**Exercise 2 Filtering Streams (8 + 8 points)**

Text

Description automatically generated

Graphical user interface, text, email

Description automatically generated

In general, the probability of a false positive is the probability of a 1-bit raised to the power of k, I, e :

* 1. m=2bil, n=10bil, k=3= (1-0.54881)^3 = 0.4512^3 = 0.09185
  2. m=2bil, n=10bil, k=4= (1-0.44933)^4 = 0.55067^4 = 0.09195

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| m=2 n=10 | K=3 | K=4 | K=5 | K=6 |
|  | 0.0918487 | 0.09195348 | 0.10092499 | 0.11644961 |

2

basing on the calculation, we can have a chart below:

Chart, line chart

Description automatically generated

When there are 3 hash functions, A minimal false positive rate is appeared.

## Exercise 3 PageRank (22+13 points)

http://snap.stanford.edu/data/web-Google.htmlText, letter

Description automatically generated

## The top 10 highest rank of the pages:

|  |  |  |
| --- | --- | --- |
| No | Node ID | Rank |
| 1 | 163075 | 6.57930344155782E-4 |
| 2 | 597621 | 6.538465907537884E-4 |
| 3 | 537039 | 6.41168046820425E-4 |
| 4 | 41909 | 6.250066824853455E-4 |
| 5 | 384666 | 5.398350690554848E-4 |
| 6 | 504140 | 5.375654569547856E-4 |
| 7 | 605856 | 5.267175479252007E-4 |
| 8 | 551829 | 5.168942917752945E-4 |
| 9 | 486980 | 5.135998255216842E-4 |
| 10 | 558791 | 5.046655488181088E-4 |

The codes and the relative document are in the zip.

## Exercise 4 Data streams (7 + 7 points)

Text, letter

Description automatically generated

## 4.1 Input stream = 3, 1, 4, 6, 5, 9

Q1: h(x) = (2x+1) mod 32

h(3) = 7 = 00111 = 0

h(1) = 3 = 00011= 0

h(4) = 9 = 01001= 0

h(6) = 13 = 01101= 0

h(5) = 11 = 01011= 0

h(9) = 19 = 10011= 0

Max Tail Length R = 0

Estimate number of distinct elements 2^R= 2^0 = 1

Q2: h(x) = (3x+7) mod 32

h(3) = 16 = 10000=4

h(1) = 10 = 11010=1

h(4) = 19 = 10011=0

h(6) = 25 = 11001=0

h(5) = 22 = 10110=1

h(9) = 2 = 00010=1

Max Tail Length R = 4

Estimate number of distinct elements 2^R = 2^4 =16

Q2: h(x) = 4x mod 32

h(3) = 12 = 10000=4

h(1) = 4 = 00100=2

h(4) = 16 = 10000=4

h(6) = 24 = 11000=3

h(5) = 20 = 10100=2

h(9) = 4 = 00100=2

Max Tail Length R = 4

Estimate number of distinct elements 2^R = 2^4 = 16

Text

Description automatically generated

Q4: h(x) = (6x+2) mod 32

h(4) = 26 = 11010 = 1

h(5) = 0 = 00000= 0

h(6) = 6 = 00110= 1

h(7) = 12 = 01100= 2

h(10) = 30 = 11110= 1

h(15) = 28 = 11100= 2

Max Tail Length R = 2

Estimate number of distinct elements 2^R= 2^2 = 4

Q5: h(x) = (2x+5) mod 32

h(4) = 13 = 01101 = 0

h(5) = 15 = 01111= 0

h(6) = 17 = 10001= 0

h(7) = 19 = 10011= 0

h(10) = 25 = 11001= 0

h(15) = 3 = 00011= 0

Max Tail Length R = 0

Estimate number of distinct elements 2^0 = 2^0 =1

Q6: h(x) = 2x mod 32

h(4) = 8 = 01000 = 3

h(5) = 10 = 01010= 1

h(6) = 12 = 01100= 2

h(7) = 14 = 01110= 1

h(10) = 20 = 10100= 2

h(15) = 30 = 11110= 1

Max Tail Length R = 3

Estimate number of distinct elements 2^R = 2^3 = 8

## Exercise 5 Summary of 3.6 and 3.7 (10 +10 points)

Summarize the content of 3.6 in your own words (600 words). (Emily)

Summarize the content of 3.7 in your own words (600 words). (Jin)