Bloom Filters

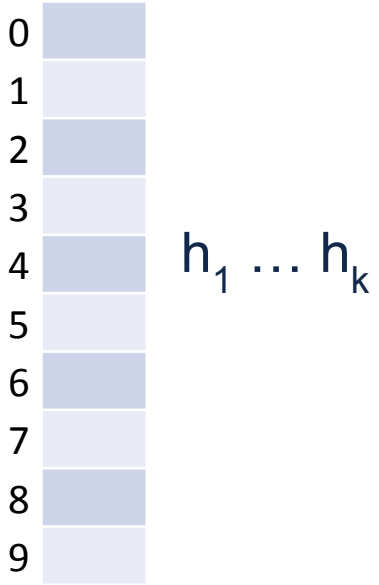
Learning Objectives

1. Know the Bloom Filter False Positive Rate

Deriving False Positive Rate

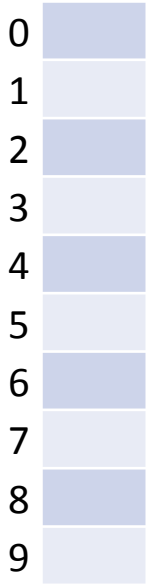
Given a bit vector of length m with k SUHA hash functions and n items inserted,

What is the expected False Positive Rate (FPR)?



Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?

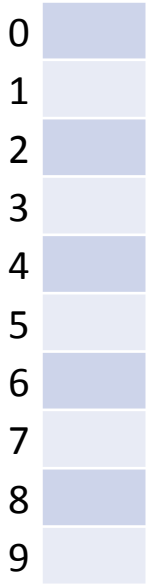


Given item x, not in the dataset, when looking for x in the bloom filter,
what is the probability that all hash values $h_1(x) \dots h_k(x)$
are all 1 in respective Bloom Filters?

$h_1 \dots h_k$

Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?



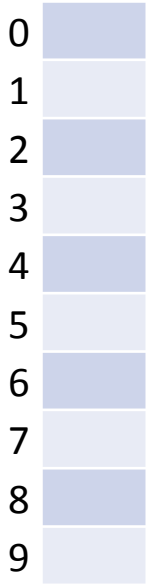
What's the probability a specific bucket is 1 after one item is inserted?

$h_1 \ldots h_k$

What about after n items?

Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?



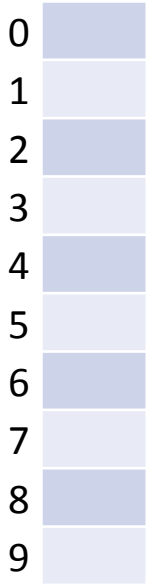
What's the probability a specific bucket is 0 after one item is inserted?

$h_1 \ldots h_k$

What about after n items?

Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?

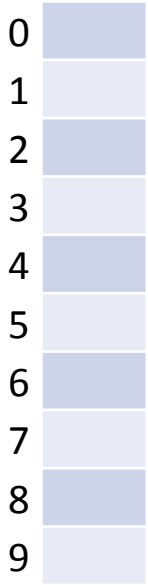


What about after k hash functions?

$h_1 \ldots h_k$

Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?

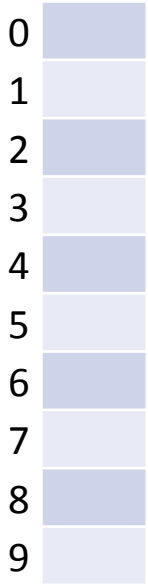


Probability that k buckets are all 0 after n insertions

$h_1 \ldots h_k$

Deriving False Positive Rate

Given a bit vector of length m with k SUHA hash functions and n items inserted,
What is the expected False Positive Rate (FPR)?

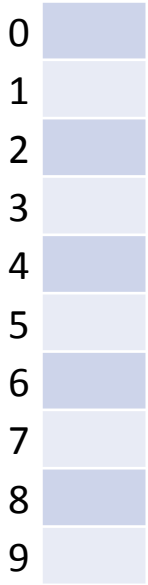


When do I get a False Positive?

$h_1 \ldots h_k$

Deriving False Positive Rate

The probability my bit is 1 after n objects inserted



$h_1 \ldots h_k$

$$\left(1 - \left(1 - \frac{1}{m}\right)^{nk}\right)^{k}$$

The number of checks for the new value against [assumed independent] trials

Minimizing False Positive Rate

Do I want high or low:

- m
- n
- k

$$\left(1-\left(1-\frac{1}{m}\right)^{nk}\right)^{k}$$

Minimizing False Positive Rate

Larger k requires more values to be 1 and raises the error rate

Larger k also provides more independent tests which helps lower the error rate

$$\left(1 - \left(1 - \frac{1}{m}\right)^{nk}\right)^{k}$$