BF = Bloom filters

set S of n keys : membership $x \stackrel{?}{\in} S$ 1-side error with pr. f

$f \sim (1-p)^k$ failure probability best $p = \frac{1}{2}$ $\Rightarrow$ $f \sim \frac{1}{2^{\frac{m}{n} \ln 2}} \sim 0.618^{m/n}$

$k = \frac{m}{n} \ln 2$ # hash functions

space : m bits + $O(k)$ words

m chosen to get a certain f

More info on the space:

$\log_2 f \sim \frac{m}{n} \ln 2 \cdot \underbrace{\log_2 \frac{1}{2}}_{-1} \Rightarrow \boxed{m \sim \frac{n \log f}{\log \frac{1}{2} \ln 2} = \frac{n \log 1/f}{\ln 2} \sim 1.44 \cdot n \log_2 1/f}$

bits

Approximate dictionary

- set S of n keys
- $x \in S$ with error probability ε

BF is one possible answer

Is it the best possible?

$O(k)$ time

space $\sim 1.44 \log_2 \frac{1}{\varepsilon}$ bits per key

① LOWER BOUNDS

- TRIVIAL: $\Omega(1)$ time
- $\geq n \log_2 \frac{1}{\varepsilon}$ bits

② UPPER BOUNDS

- $O(1)$ time with 1 hash function
- $\sim n \log_2 \frac{1}{\varepsilon}$ bits + lower order terms

Information theory lower bound

$S \subseteq U$ and let $|U| = m$ ($\neq m$ in BF!)
$|S| = n$

Q. How many sets $S \subseteq U$ of size n? $\binom{m}{n}$ choices

I.T. says that using less than $\lg_2 \binom{m}{n}$ in the worst case cannot give a correct algorithm: indeed, using less bits forces two S' and S'' to get the same binary representation so membership is fooled

exact

A useful formula:

$$\lg_2 \binom{m}{n} \sim n \lg_2 \frac{m}{n}$$

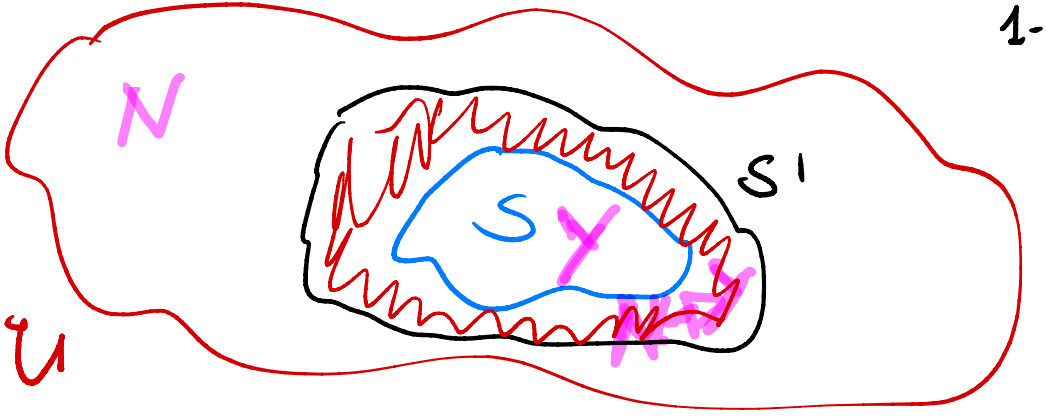
Approximate dictionary for S with pr. $\frac{1}{f}$, D = exact dictionary for S

D' :

Exact dictionary for some $S' \supseteq S$ and $\frac{|S' \setminus S|}{|U|} = \frac{1}{f}$

$S' = \{x \in U : \text{approximate dictionary says YES}\}$

$S' = \{x \in U : D'(x) = \text{True}\}$



1-side error for $S \Rightarrow$

$S \subseteq S'$ all keys in S are accepted plus the extra (wrong) keys in S'/S

Q: Let D' be the exact dictionary for S'

D " " " S

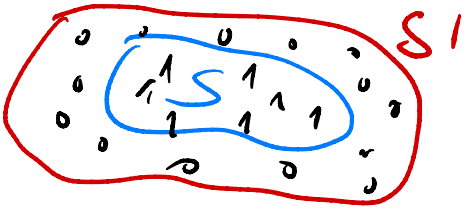
(note that D' is also the approximate dictionary for S)

➡ Given D', can we get D?

b' = #bits required by D'

$D = D' +$ extra bits to mark correct answers from D'

we have $|S'|$ YES answers, and we mark the i-th YES answer from D' $< \begin{matrix} 1 & x \in S \\ 0 & x \notin S \end{matrix}$ (have $x \in S' \setminus S$)



We get an exact dictionary D $\Rightarrow$ it must require $\log_2 \binom{m}{n}$ bits by I.T.

$$|D| \geq \log_2 \binom{m}{n} \quad \text{I.T.}$$

$$|D' + \text{extra bits}| \geq \log_2 \binom{m}{n}$$ as D' + extra bits are a dictionary for S

$$|D'| + \log_2 \binom{|S'|}{|S|} = b' + \log_2 \binom{n+fm}{n}$$

$$(|S'| = |S| + |S'/S| = |S| + |U| \cdot f = n + fm)$$

$$\frac{|S'/S|}{|U|} = f \Leftrightarrow |S'/S| = fm$$

Putting these parts together:

$$b' + \log\binom{n+fm}{n} \geq \log\binom{m}{n}$$

$$b' \geq \log\binom{m}{n} - \log\binom{n+fm}{n} \sim n\log\frac{m}{n} - n\log\frac{n+fm}{n} \sim n\log\frac{m}{n} - n\log\frac{fm}{n}$$

$$\sim \log\frac{fm}{n} \quad \text{as} \quad \frac{n+fm}{n} = 1+\frac{fm}{n}$$

$$= n\left(\log\frac{m}{n} - \log\frac{fm}{n}\right)$$

$$= n\left(\log\frac{m}{n} + \log\frac{n}{fm}\right)$$

$$= n\left(\log\frac{m}{n}\cdot\frac{n}{fm}\right) = n\log_2\frac{1}{f}$$

D' requires at least $n\log_2\frac{1}{f}$

UPPER BOUND

Bit vector m bits n 1s

[RRR] $\lg \binom{m}{n}$ + lower order terms

B(i) in O(1) time

constant-time lookup
succinct data structure

(1)

Universal Hashy $h \in)\{$

$\frac{1}{f} = \frac{m}{n} \Rightarrow m = \frac{n}{f}$

(2)

$\forall x \in S: \quad B[h(x)] = 1$ (BF with 1 hash function + RRR)

- Space is optimal because of RRR
- time is $O(1)$ as we use 1 hash function
- error probability : union bound

$$Pr[B[i]=1] = \sum_{x \in S} \underbrace{Pr[h(x)=i]}_{\frac{1}{m}} = \frac{n}{m} = f$$

m = universe $= |U| = \frac{n}{f}$ you store the hash values in B!