

Birthday paradox

$$\Pr[\text{2 people share the same bday}] = 1 - \frac{365}{365^2} = 1 - \frac{1}{365}$$

person by person

$$\Pr[\text{no 2 people share the same bday}] = \left(1 - \frac{1}{365}\right) \left(1 - \frac{2}{365}\right) \dots \left(1 - \frac{n-1}{365}\right)$$

$$p(2 \text{ no bd}) p(3 \text{ no bd} | 2 \text{ no bd}) p(4 \text{ no bd} | 3 \text{ no bd}) \dots p(n \text{ no bd} | n-1 \text{ no bd})$$

$$1 - \left(\frac{365}{365^2} + \frac{368}{365^2}\right) = 1 - \frac{2}{365}$$

$$[365 \cdot 364 \cdot \dots \cdot (365 - n)]$$

$$n = 60 \sim 0.5\%$$

group

$$\Pr[\dots] = \frac{\binom{365}{n} n!}{365^n} \quad \left. \begin{array}{l} \text{orderings of } \blacksquare \text{ days} \\ \text{not the same, across } n \text{ slots} \end{array} \right\}$$

$$\left. \begin{array}{l} \text{all orderings of } \blacksquare \text{ days} \\ \text{across } n \text{ slots} \end{array} \right\}$$

$$n = 60$$

$$\Pr[\dots] \sim 0.5\%$$

Generalized  $\leftarrow$  balls  $\leftarrow$  bins

$n$  "people",  $m$  "days"

$$\Pr[\text{no collision}] = \prod_{j=1}^{n-1} \left(1 - \frac{j}{m}\right) \approx \prod_{j=1}^{n-1} e^{-j/m} =$$

$$\text{Approx: } 1 - \frac{k}{m} \approx e^{-k/m}, \text{ small } k$$

$$= \exp\left[-\sum_{j=1}^{n-1} \frac{j}{m}\right] = \exp\left[-\frac{1}{m} \frac{n-1}{2} \cdot n\right] \approx e^{-n^2/2m}$$

Q: What is the value of  $n$  for which  $\Pr[\text{collision}] = 1/2$

$$\ln\left(e^{-n^2/2 \cdot 365}\right) = \ln\left(\frac{1}{2}\right)$$

$$\frac{-n^2}{2 \cdot 365} = \ln\left(\frac{1}{2}\right)$$

$$-\ln\left(\frac{1}{2}\right) = \ln(1) - \ln\left(\frac{1}{2}\right) =$$

$$= \ln\left(\frac{1}{1/2}\right) = \ln(2)$$

~~XXXXXXXXXX~~

$$n \approx 22.49$$



Q: How many bins ("days") are empty?

$$\Pr[\text{a bin is empty}] = \left(1 - \frac{1}{m}\right)^n \text{ all balls } \approx e^{-n/m} =$$

$$\mathbb{E}\left[\sum_{i=1}^m x_i\right] = \sum_{i=1}^m \mathbb{E}(x_i)$$

linearity of expectation over indicator r.v.s

$$\# \text{ empty bins} = m \cdot e^{-n/m}$$



$$= e^{-\frac{n}{m}}$$

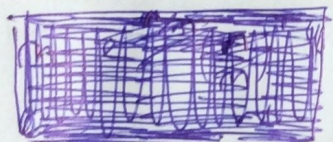
Q: prob. a bin has  $r$  balls

$$\Pr[\text{given bin has } r \text{ balls}] = \binom{n}{r} \left(\frac{1}{m}\right)^r \left(\frac{m-1}{m}\right)^{n-r}$$



$$\approx e^{-\frac{n}{m}} \frac{\left(\frac{n}{m}\right)^r}{r!}$$

Q: # of bins with  $r$  balls in each



$$m \cdot e^{-n/m} \frac{\left(\frac{n}{m}\right)^r}{r!}$$

Hash function: maps a larger set to a smaller set  
(birthday is a hash function)  
→ eventually get collisions

Defn:  $f: \{0, \dots, n-1\} \rightarrow \{0, \dots, m-1\}$



(usually  $m \ll n$ )

mapping is deterministic  
each elt in  $\{0, \dots, n-1\}$  is equally likely  
mapped to any elt in  $\{0, \dots, m-1\}$

Passwords

0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 m bits

dictionary  
of weak  
pwd's  
map to  
m bits

collision  
0 0 1 1 0 1 1 0 0 1 0 0 0 0

then check each new pwd against the array  
→ false positives due to collisions  
→ no false negatives



(2)

trade-off btwn. space and false positive rate

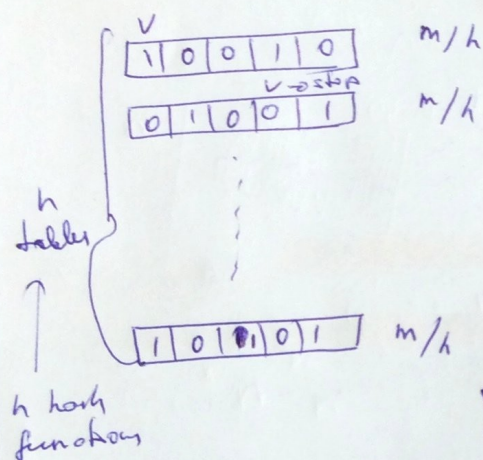
fraction  $e^{-n/m}$  bins will be empty (on average)

$$P[\text{reject strong pwd}] = 1 - e^{-n/m} \leftarrow \text{FP rate}$$

$p = e^{-n/m}$  a strong pwd hits any bin equally likely. with  $1-p$  the bin that was hit is non-empty

Bloom filters  
Smart way to deal with space vs. FP rate trade-off

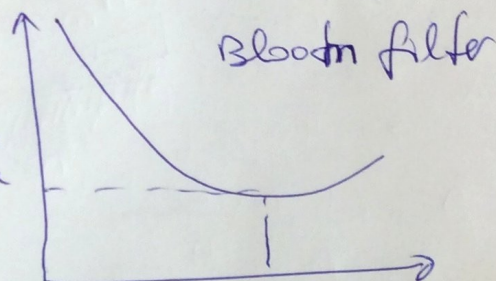
$m$  bits, but  $h$  tables, each of size  $\frac{m}{h}$



$$P_r[\text{reject a strong pwd}] = (1 - e^{-n/(m/h)})^h = (1 - e^{-nh/m})^h$$

optimal  $h$ ?

makes smaller  $\downarrow$   
makes larger  $\uparrow$



e.g. 100,000 weak pwds

$\sim 7$  chars long

700,000 bytes  $\xrightarrow{\text{compression}}$  350,000 bytes

Bloom filter: 5 tables, 160k bits  
FP = 2%

$$h_{\text{opt}} = (\ln 2) \frac{m}{n}$$