Hashing 1

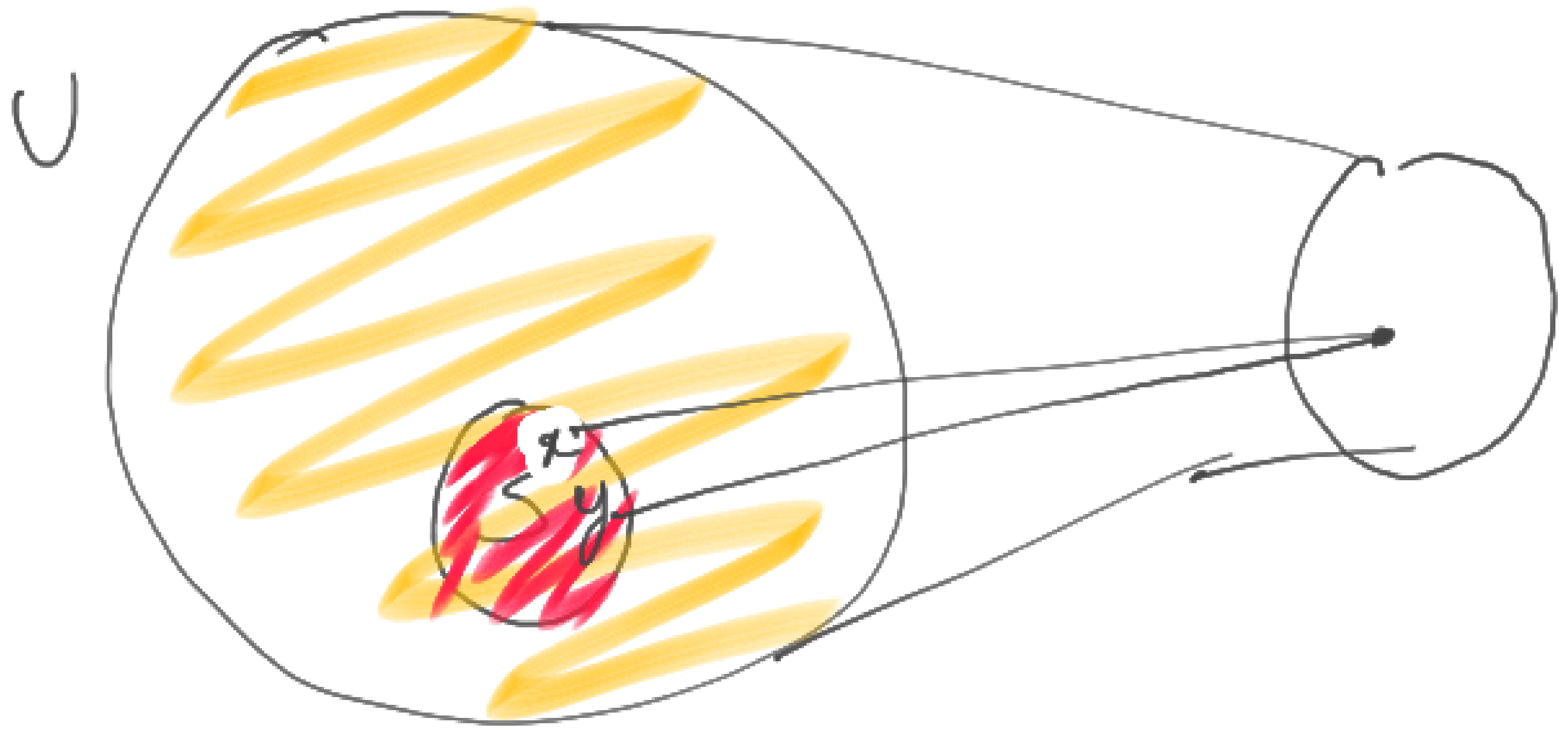
Jan 25 2022

Setting:

Universe U = set of all possible values

subset S = interested in this subset

$$|S| = N$$



w.h.p elements from S have small number of collisions

$x, y \in S$

Collision $\Rightarrow h(x) = h(y)$

w.h.p $\Rightarrow$ what is random?

① inputs

② hash function

E.g. Ntk switch : IP address x.x.x.x

Assume a family of hash functions H

When time to hash "S", we choose a random function

$h \in H$

M = table size

$[M] = \{0, 1, \ldots, M-1\}$

What properties ?

1. Small probability of distinct keys colliding :

   If $x \neq y$, $P[h(x) = h(y)]$ is 'small'

2. h is easy to compute
3. h is easy to store (small num. of bits)
4. M is small $\Rightarrow$ (hash table size is small)

Opposing

Ideal Hash Function

Perfectly random:
for each $x \in S$ $h(x) =$ a uniformly random location in $[M]$

Properties:

- Low collision prob. $P[h(x) = h(y)] = \frac{1}{M}$ for any $x \neq y$
- Even conditioned on hashed values for any other subset A of S

Downsides:

1. Too large to store
2. Compute : table look up

Universal Hash Functions

Captures the property of non-collision of two distinct elements.

Dfn: A family $\mathcal{H}$ of hash functions mapping $U \to [M]$ is universal if for any $x \neq y$

$$P\,[h(x) = h(y)] \leq \frac{1}{M}$$

needs to hold for every pair x & y, $x \neq y$

Simple construction of universal hashing:

Let $|U| = 2^u$ elements are (binary vectors of length "u"

$|M| = 2^m$ ______ " ______ "m"

Let $\underset{m \times u}{A}$ → binary matrix (uniform random binary elements)

$$\underset{m\times 1}{h(x)} = \underset{m \times u}{\boxed{A}}\ \underset{u \times 1}{\boxed{x}}$$

For any $u \in U$, $\underset{m\times 1}{h(x)} := \underset{m\times u}{A}\ \underset{u\times 1}{x}$

(modulo 2 arithmetic)

Q: How many hash fns?

2^{um}

Thm: This family of hash function is <u>universal</u>.

To show

$$P[h(x)=h(y)] \le \frac{1}{M}$$

for $x \neq y$

Proof:

$$h(x) = h(y) \quad \text{where } x \neq y$$

$$Ax = Ay$$

$$A(x-y) = 0$$

$$Az = 0 \quad \text{for } z \neq 0$$

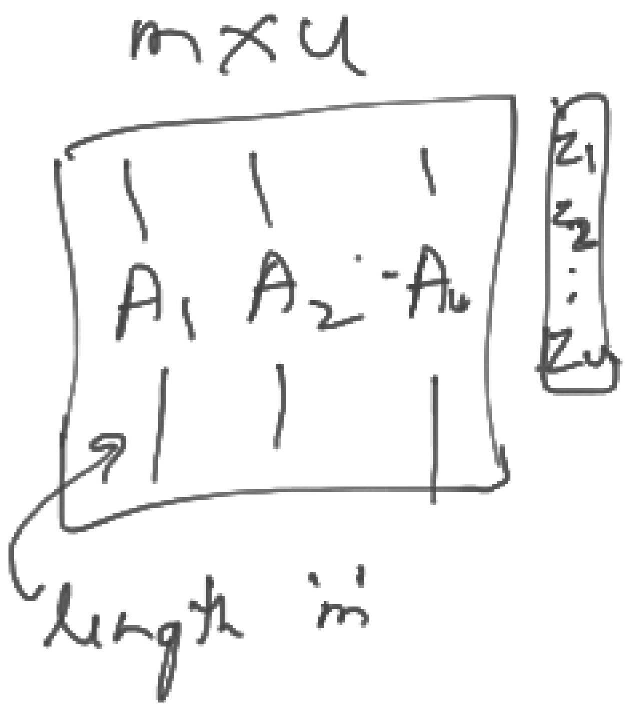
! Why? $\because x \neq y$

To show:

$$P(Az = 0) \le \frac{1}{M} \quad \text{for any } z \neq 0$$

$$Az = 0 \Rightarrow \sum A_j z_j = 0$$

columns of A



Let $z_{i^*} \neq 0$ $\quad (\exists_j \, i^* : z \neq 0)$

there exists

a

$(\Rightarrow z_{i^*} = 1)$

$Az = 0$

$$\Rightarrow \quad A_{i^*} = - \sum_{j \neq i^*} A_j z_j$$

fixed binary vector

e.g.

$$p(Az = 0) = p\left(A_{i^*} = - \sum_{j \neq i^*} A_j z_j\right)$$

$$= \left(\frac{1}{2}\right)^m$$

$$= \frac{1}{M}$$

$$Az = \sum A_j z_j$$

$$Az = 0$$

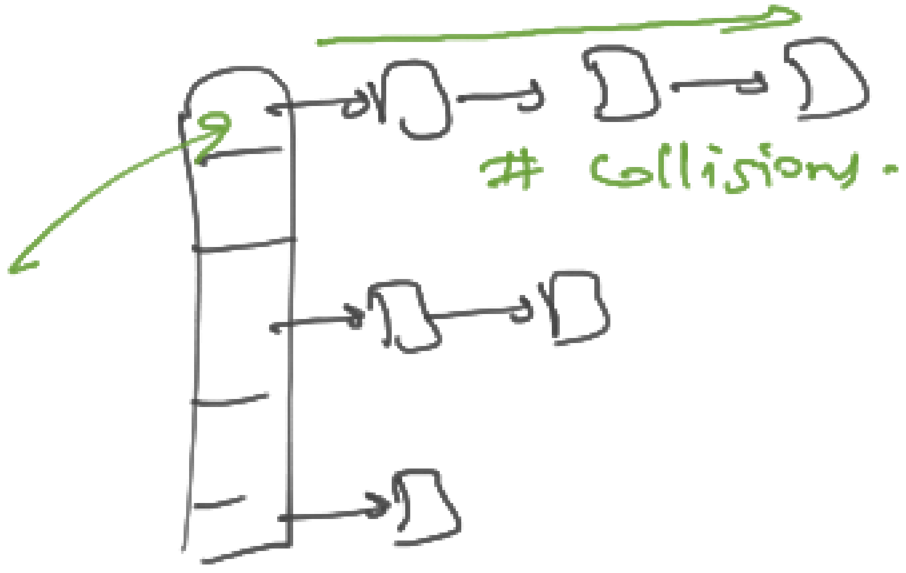
$$\Rightarrow \sum_{j=1}^{u} A_j z_j = 0$$

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} = \begin{matrix} A_{i^*0} = 0 \\ A_{i^*1} = 1 \\ \vdots \\ A_{i^*m-1} = 0 \end{matrix}$$

$$A_1 z_1 + A_2 z_2 + \dots A_u z_u = 0$$

$$A_{i^*} z_{i^*} = - \sum_{j \neq i^*} A_j z_j$$

Application 1: Hash tables

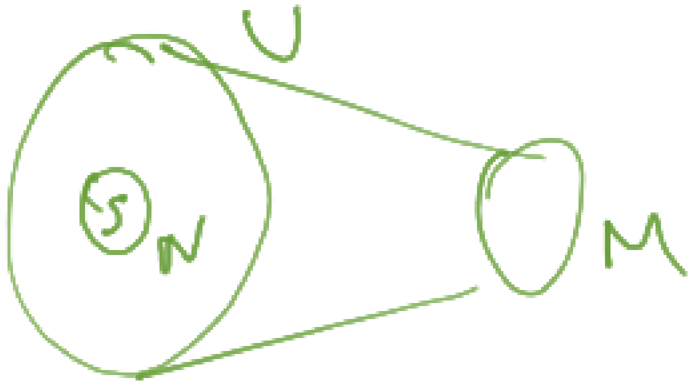


Closed addressing

Open addressing

Look up time : number of collisions.

Let C_x = num. of other elements mapped to the value where x is mapped to.



$$L_x = C_x + 1$$

$$E[L_x] = \underset{?}{E[C_x]} + 1 = \frac{(N-1)}{M} + 1$$

If $M \geq N$ $\quad E[L_x] \leq 2$! in expectation

Let C = total number of collision.

$$E[C] \le \binom{N}{2} \cdot \frac{1}{M}$$

$$\frac{N(N-1)}{2}$$

Collision free hash table?

If $M \geq N^2$

P[there exists a collision] = $\frac{1}{2}$

Repeat the exp to get collision-free $\Rightarrow$ Const. look up time in worst-case

Downside: $M \geq N^2$

Q. Can we get collision free with $M = O(N)$