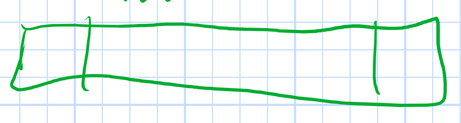
L14: Hashing

10/10/23

Birthday paradox

$s_1, s_2, \dots, s_n$ n items

throw them into



m buckets

$Z = \#$ of colliding pairs

$1 \le i < j \le n$

$X_{i,j} = 1 \iff s_i$ and s_j are in the same bucket.

$$P[X_{i,j}=1] = \frac{1}{m}$$

$$E[Z] = E\Big[\sum_{i<j} X_{i,j}\Big] = \sum E[X_{i,j}] = \sum_{i<j} P(X_{i,j}=1) = \binom{n}{2}\frac{1}{m}.$$

$m = n$

$$E[Z(n,n)] = \frac{1}{n}\binom{n}{2} = \frac{1}{n}\frac{n(n-1)}{2} \le \frac{n}{2}$$

$m = n^2$

$$E[Z(n,n^2)] = \frac{1}{n^2}\binom{n}{2} = \frac{n(n-1)}{2n^2} \le \frac{1}{2}.$$

$$P\left[\begin{array}{l}\text{there is a collusion}\\ \text{when throwing } n \text{ items into}\\ n^2 \text{ b slots}\end{array}\right] = P[Z \ge 1] \underset{\text{Markov}}{\le} \frac{E[Z]}{1} = \frac{1}{2}.$$

---

h: hash function

$h: U \to [m] = \{1, 2, \dots, m\}$

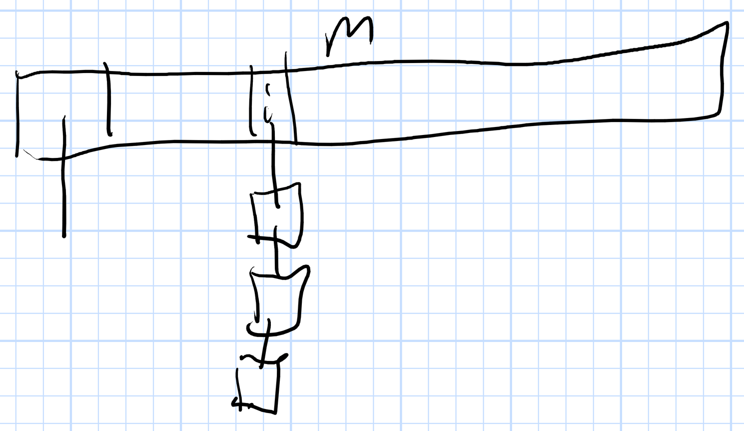
$U \supseteq S$

$\log m^n = \boxed{n \log n}$

$H \leftarrow$ set of functions, randomly pick one of them

$$\boxed{P_{h\in H}[h(x)=h(y)] = \frac{1}{m} \longleftarrow H \text{ is 2-universal if}}$$

Open hashing.



m

Lemma

If H is 2-universal then using open hashing for n items and array of size m then the expected time per operation is $O(1+\frac{n}{m})$.

Proof Let x be the item we operate on.

$$E\left[\begin{array}{l}\#\text{ of colliding}\\ \text{items with } x\end{array}\right] = \sum_{i=1}^{n} 1 \cdot P\left[\begin{array}{l}i\text{th item}\\ \text{collide with } x\end{array}\right] = n\frac{1}{m} = \frac{n}{m}.$$

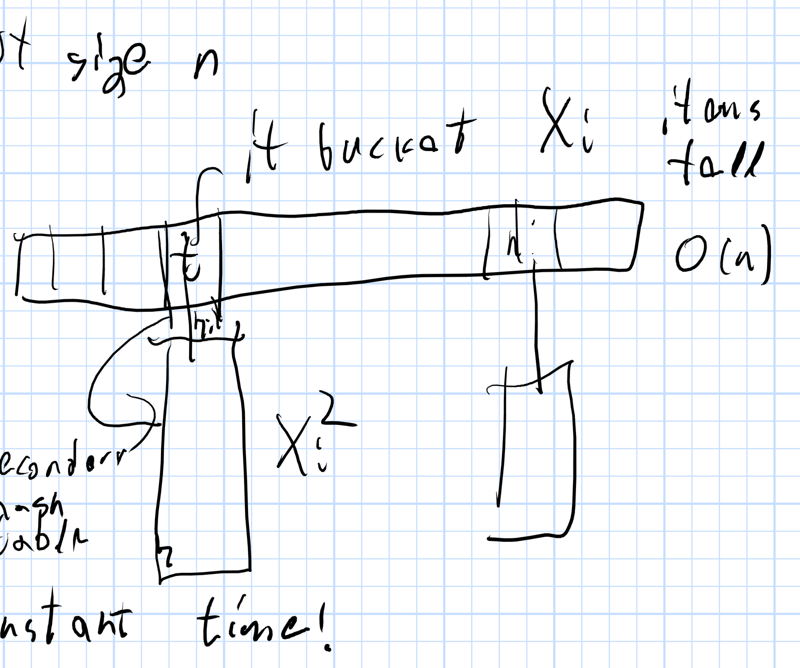
$O(1+\frac{n}{m})$ expected time per operation. ∎

Perfect hashing

$S = \{s_1, s_2, \dots, s_n\}$

$S \xrightarrow{} $ Array of size n

↑ 2-Universal hash function



Lookup in constant time!

Analyzing the space

$X_i \equiv \#$ of items falling into the ith bucket

$\sum X_i = n$

$$X_i^2 = 2\frac{X_i(X_i-1)}{2} + X_i = \overbrace{2\binom{X_i}{2} + X_i}^{=X_i^2}$$

total space is $\underbrace{\sum X_i^2 + n}$

$$E[\text{space}] = E\Big[n + \sum X_i^2\Big] = n + \sum_{i=1}^{n} E[X_i^2] = n + 2\sum E\Big[\binom{X_i}{2}\Big] + \overbrace{\sum_{i=1}^{n} E[X_i]}^{n}$$

$$= 2n + 2\sum_{i=1}^{n} E\Big[\binom{X_i}{2}\Big] \le 2n + 2\cdot\frac{1}{2}n = 3n$$

$\binom{X_i}{2} = \#$ of pairs of items colliding in the ith bucket

$\sum_{i=1}^{n}\binom{X_i}{2} =$ total number of colliding pairs

$$E[\ \ ] = E[\ \ ] = \binom{n}{2}\frac{1}{m} \overset{m=n}{=} \frac{n(n-1)}{2n} \le \frac{n}{2}$$

Data structure is BAD if space used is larger than $6n$

$$P\left[\begin{array}{l}\text{algorithm fails}\\ \text{in picking}\\ \text{top level hash function}\end{array}\right] = P[\text{space} \ge 6n] \overset{\text{Markov}}{\le} \frac{E[\text{space}]}{6n} = \frac{3n}{6n} = \frac{1}{2}.$$

Theorem

Given 2-universal hash functions, one can construct a perfect hashing scheme for n items (provided in advance) that uses $O(n)$ space, and lookup takes $O(1)$ time.

---

p a sufficiently large prime number.

How many prime numbers in $\{1, 2, \dots, n\}$

The number of primes $[n] \approx \frac{n}{\log n}$

$\mathbb{Z}_p = \{0, 1, 2, \dots, p-1\}$

$i \pmod p =$ remainder of dividing i by p.

↑ integer

Theorem $\mathbb{Z}_p$ is a field.

S set of items

$S \subseteq U$ $|U| < p$ m target array size

$H = U \to \{0, \dots, m-1\}$

$H = \{(ax+b \pmod p) \bmod m \mid a, b \in \mathbb{Z}_p\ \ a \ne 0\}$

$((ax+b) \bmod p) \bmod m$