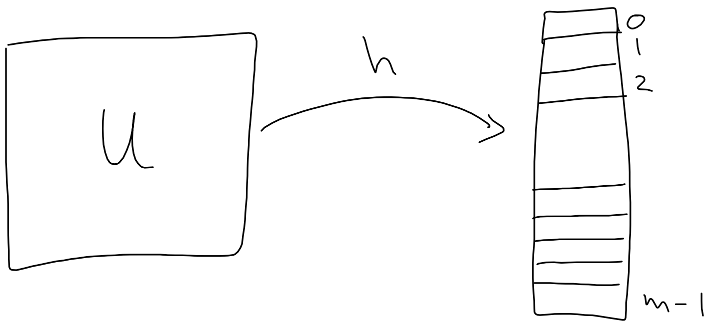
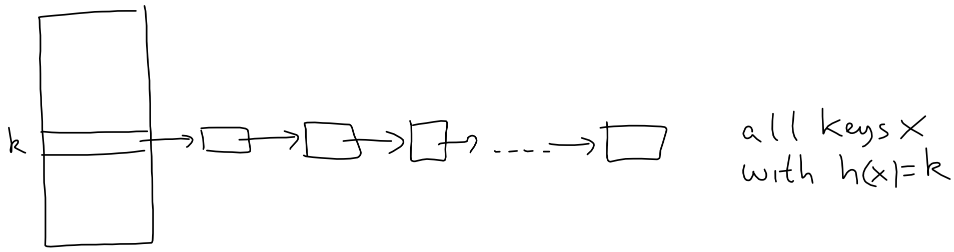
Hashing





Let $\mathcal{H}$ be a finite collection of hash-fcts

$U: \longrightarrow \{0,1,2,\dots,m-1\}$

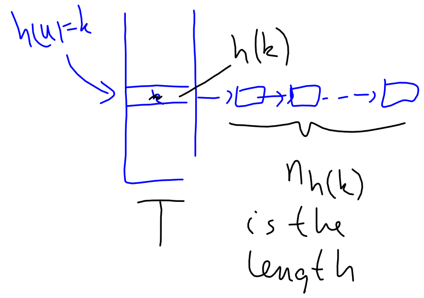
$\mathcal{H}$ is <u>universal</u> if :

$$\forall k,l \in U : \left|\{h \in \mathcal{H} \mid h(k)=h(l)\}\right| \leq \frac{|\mathcal{H}|}{m}$$

so if we choose a random $h \in \mathcal{H}$ we have $\text{prob}(h(k)=h(l)) \leq \frac{1}{m}$

Thm 11.3 Suppose $\mathcal{H}$ is a universal family of hash fcts $U \to \{0,1,\dots,m-1\}: T$

hash n elem from U using random $h \in \mathcal{H}$ and we use chaining for collisions



- If $k \notin T$ then

$$E(n_{h(k)}) \leq \frac{n}{m} = \alpha$$

- If $k \in T$ then

$$E(n_{h(k)}) \leq 1 + \alpha$$

P: Define for each i,j $X_{i,j} = \begin{cases} 1 & \text{if } h(i)=h(j) \\ 0 \end{cases}$

Y_k = # keys with $h(k)=h(\ell)$ and $\ell \neq k$

Then $E(Y_k) = E\left(\sum_{\substack{\ell \neq k \\ \ell \text{ is in table}}} X_{k\ell}\right)$ $\left(Y_k = \sum_{\substack{\ell \neq k \\ \ell \in T}} X_{k\ell} \right)$

if $k \notin T$ $\quad E(Y_k) = E(n_{h(k)})$

$$= \sum_{\substack{\ell \neq k \\ \ell \text{ in } T}} E(X_{k\ell}) = \sum_{\ell \neq k, \ell \text{ in } T} p(h(k)=h(\ell))$$

$$\leq \sum_{\ell \neq k} \frac{1}{m} = \frac{n}{m}$$

If $k \in T$ then $E(Y_k) = \sum_{\ell \neq k, \ell \in T} \frac{1}{m} = \frac{n-1}{m}$

So $E(n_{h(k)}) = 1 + E(Y_k) \leq 1 + \frac{n-1}{m}$

How do we obtain/construct a universal family $\mathcal{H}$?

$a = bk \bmod p$ $\quad$ p prime

$\Downarrow ak^{-1} = b$ where $k \cdot k^{-1} \equiv 1 \bmod p$

(ex $p = 7$: $\quad 6 \cdot 6 = 36 = 1 \bmod 7$

so $6^{-1} = 6$)

need to know:

if p is a prime then we can solve

$a = bk \bmod p$ and find b.

- Choose prime p such that $U \subseteq \{0,1,2,\dots p-1\}$

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

- p prime $\Rightarrow$ we can solve equations modulo p

- $p \geq |U| > m$ so $p > m$

- define $h_{a,b}$ $\quad \forall a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p$

  by $h_{a,b}(x) = ((ax+b) \bmod p) \bmod m$

- $\mathcal{H} = \mathcal{H}_{p,m} = \{ h_{a,b} \mid a \in \mathbb{Z}_p^*, b \in \mathbb{Z}_p \}$

  $|\mathcal{H}_{pm}| = (p-1)p$

Thm 11.5 $\mathcal{H}_{pm}$ is universal

(so we claim that if $h_{a,b} \in \mathcal{H}_{pm}$ is randomly chosen, then
$\forall k, \ell \in \mathbb{Z}_p : p(h_{a,b}(k) = h_{a,b}(\ell)) \leq \frac{1}{m}$)

P: fix $h_{a,b}$ (random from $\mathcal{H}_{pm}$) and $k, \ell \in \mathbb{Z}_p$ $k \neq \ell$

let (*) $r = ak + b \bmod p$ $\quad s = a\ell + b \bmod p$

so $r - s = a(k - \ell) \bmod p \neq 0$ since p is a prime and $a, (k-\ell) \leq p-1$

So no collisions at mod p level!

($\square$) $\boxed{\text{There are } p(p-1) \text{ choices for } a,b}$

recall $(*)$ $r = ak+b \bmod p$ $s = a\ell+b \bmod p$

solve for a,b in $(*)$

$r-s = a(k-\ell) \bmod p$ $b = r - ak \bmod p$

$(r-s)(k-\ell)^{-1} = a \bmod p$

so we k,ℓ $(k \neq \ell)$ fixed then is precisely one solution a,b to $(*)$

($\square$)+ only $p(p-1)$ possible pairs r,s with $r \neq s$

1-1 correspondance between pairs (a,b) and pairs r,s

Conclusion: Given $k \neq l$ as input

if we pick a, b randomly $a \in \mathbb{Z}_p^\times, b \in \mathbb{Z}_p$

the resulting (r, s) is equally likely to be

any pair of distinct values mod p

Thus $p(h_{a,b}(k) = h_{a,b}(l))$ = prob that $r \equiv s \mod m$

for random r, s $r \neq s$

For fixed $r \in \mathbb{Z}_p$ there are $p-1$

remaining values for $s \neq r$ and at most

$\left\lceil \frac{p}{m} \right\rceil - 1$ of these will have $s \equiv r \mod m$

comes from r

$$\left\lceil \frac{p}{m} \right\rceil - 1 \leq \frac{p+m-1}{m} - 1 = \frac{p-1}{m}$$

So prob that $s \equiv r \mod m$ is $\leq \frac{1}{p-1} \cdot \frac{p-1}{m} = \frac{1}{m}$

Perfect hashing (using univ. hashing)

How large a hash table T $m=|T|$ do we need in order to avoid (expect no) collisions?

- with n distinct elements there are $\binom{n}{2}$ possible collisions

Lemma: If $m=n^2$ then $E(X) < \frac{1}{2}$ and $p(X \geq 1) < \frac{1}{2}$

where $X = \#$ collisions

P: $X_{ij} = \begin{cases} 1 & \text{if } h(i)=h(j) \\ 0 & \end{cases}$ $p \leq \frac{1}{m}$ univ. hash

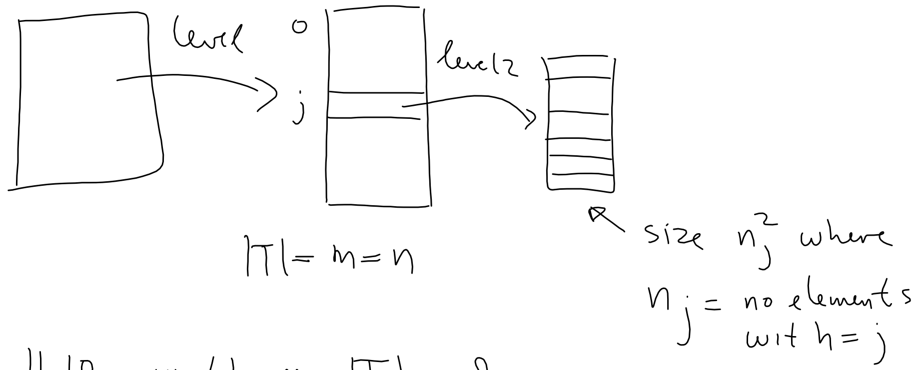
$$E(X) = E\left(\sum_{i<j} X_{ij}\right) = \sum_{i<j} p(X_{ij}=1) \leq \frac{\binom{n}{2}}{m}$$

so $E(X) \leq \frac{n(n-1)}{2n^2} = \frac{n-1}{2n} < \frac{1}{2}$

$\Rightarrow$ by Markov's ineq $p(X \geq 1) \leq \frac{E(X)}{1} < \frac{1}{2}$

$|T| = n^2$ may be far too large!

Instead use two levels of hashing



Thm 11.10 with $m = |T|$ and n_j's are above

$$E\left(\sum_{j=0}^{m-1} n_j^2\right) < 2n$$

P: use $\boxed{\forall a \in \mathbb{Z}^+ \quad a^2 = a + 2\binom{a}{2}}$

$$K = E\left(\sum_{j=0}^{m-1} n_j^2\right) = E\left(\sum_{j=0}^{m-1}\left(n_j + 2\binom{n_j}{2}\right)\right)$$

$$= \sum_{j=0}^{m-1} E(n_j) + 2\,E\left(\sum_{j=0}^{m-1}\binom{n_j}{2}\right)$$

$$= n + 2r$$

where r is expected number of collisions

[There $\binom{n}{2}$ possible collisions each happen with prob $\leq \frac{1}{m}$ so $E(\#\text{cols}) \leq \frac{\binom{n}{2}}{m}$

$$K = n + 2r$$

$$\leq n + \frac{n(n-1)}{n} = n + n - 1 < 2n$$

Corollary if we use 2 level univ hashing
with $m = n$ at level 1 and
$m_j = n_j^2$ for $j = 0, 1, 2 \ldots m-1$ at level 2

then the expected storage need is at most
$2n$ for second level (and n for first)

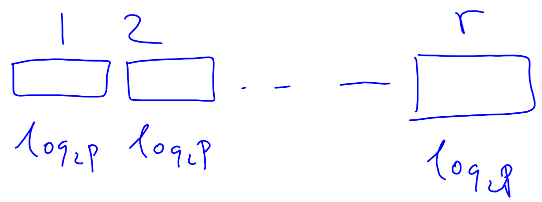
Corollary 11.12 ⇐ Markov
with hashing scheme as above
the probability that we need more
$4n$ storage for 2nd level is $< \frac{1}{2}$

Conclusion: with a few retries
we can construct a hash scheme
with no collisions

Universal hashing ala Kleinberg-Tardos

· Identify U with vectors $(x_1, x_2 .. x_r)$

$r = \dfrac{\log_2 N}{\log_2 p}$ $\quad p \approx n \quad p$ prime



U integers in $\{0,1,2 - - N-1\}$

Fix $A = \{ (a_1, a_2 - - a_r) \mid a_i \in \mathbb{Z}_p,\ i \in \{1,2,..r\} \}$

For $a \in A$ $\quad h_a(x) = \sum_{i=1}^{r} a_i x_i \mod p$

Thm $\mathcal{H} = \{ h_a \mid a \in A \}$ is universal

P: Given $x = (x_1, x_2, .. x_r)$ $\quad y = (y_1, - - y_r)$ $\quad x \neq y$

show $P(h_a(x) = h_b(y)) \leq \frac{1}{p}$ if $a \in A$ is random

If $h(x) = h(y) \mod p$

then

$$(*) \quad \sum_{i=1}^{r} a_i x_i = \sum_{i=1}^{r} a_i y_i \mod p$$

fix j s.t $x_j \neq y_j$ (as $x \neq y$ this is possible)

$(*) \iff$

$$s = \sum_{i \neq j} a_i (x_i - y_i) = a_j (y_j - x_j) \mod p$$

only one choice (out of p) for a_j

will give $=$ above

so prob of $=$ is $\leq \frac{1}{p}$