Data Structure

Data $x \in U$

Queries $f_1, f_2, \dots f_m : U \to \{0,1\}$

Goal: Preprocess x and set up memory so that when given $i \in [m]$, we can quickly compute $f_i(x)$.

Set Membership.

Set $S \subseteq U$

$|U| = M$

$|S| \le n$

Queries: For each $i \in U$, "is $i \in S$".

sets of size $\le n$ is $\binom{M}{\le n}$.

$\Rightarrow$ Any data structure for this needs space $\ge \log \binom{M}{\le n}$ bits.

$\approx O\left(n \log\left(\frac{M}{n}\right)\right)$

$\approx n \log M$. (for M big)

Word size $\approx O(\log M)$

$\Rightarrow O(n)$ words.

Going to see:

$O(n)$ words of memory used

$O(1)$ time to respond to queries.

[FKS '70s]

Pick random map $h : U \to [n^2]$

Hope: h is one-to-one on S.

Then here is a nice representation:

For each $j \in [n^2]$

Store: ① $b_j \in \{0,1\}$

$b_j = 1$ iff $\exists\, s \in S$ s.t. $h(s) = j$

else 0.

② If $b_j = 1$

store the $s \in S$ s.t. $h(s) = j$.

Pairwise independent hash family

$\mathcal{H} \subseteq \{ f : U \to [K] \}$

s.t. ① For any fixed $u \in U$

for a random $h \in \mathcal{H}$

the distribution of $h(u)$

is uniform over $[K]$.

② For any $u_1, u_2 \in U$, $u_1 \ne u_2$

for random $h \in \mathcal{H}$

the dist. $(h(u_1), h(u_2))$

is uniform over $[k] \times [k]$

$\exists$ such $\mathcal{H}$ with $|\mathcal{H}| = O(|U| \cdot k)$
$= O(M \cdot k)$

So can specify an elt of $\mathcal{H}$ using $O(\log M + \log k)$ bits.

<u>Claim</u> $\forall\, S \subseteq [M]$, $|S| = n$

$$\Pr_{h \in \mathcal{H}}\left[\forall x \neq y \in S,\ h(x) \neq h(y)\right] \geq 1 - \frac{\binom{n}{2}}{k}$$

<u>Proof</u> Fix $x, y \in S$.

$$\Pr_h\left[h(x) = h(y)\right] = \frac{1}{k}.$$

Union bound over all $x, y \in S$.

New idea:

Study the number of collisions when hashing a set of size n to a range of size $O(n)$.

For each $x, y \in S$, $(x \neq y)$ Let

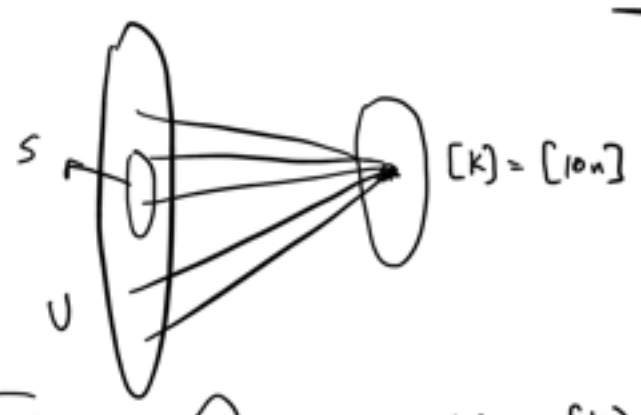
$Z_{x,y} = 1$ if $h(x) = h(y)$

$$\mathbb{E}[Z_{x,y}] = \frac{1}{k}.$$

$$Z = \sum_{\substack{x \neq y \\ x, y \in S}} Z_{x,y} = \#\text{ collisions}.$$

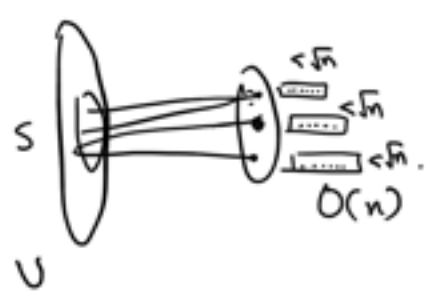
$$\mathbb{E}[Z] = \sum_{x \neq y} \mathbb{E}[Z_{x,y}] = \frac{\binom{n}{2}}{k}.$$

If $k = 10n$, Then $\mathbb{E}[Z] \leq \frac{n}{10}$.



[If $Z \leq \frac{n}{5}$, then $\forall j \in [k]$, $|h^{-1}(j) \cap S| \leq \sqrt{n}$

$$\Pr\left[Z > \frac{n}{5}\right] \leq \frac{\mathbb{E}[Z]}{n/5} \leq \frac{5}{10} \leq \frac{1}{2}.$$



Final Data Structure

- First hash from U to $[k]$ where $k = O(n)$.

Then for each $j \in [k]$, let

$$S_j = \{ s \in S : h(s) = j \} = S \cap h^{-1}(j)$$

Use the $O(n^2)$ space data structure for S_j.

Total space used $= \sum_j O(|S_j|^2) + O(n)$

$$= O(z) + O(n).$$

$$= \underline{O(n)}.$$

For each $j \in [k]$, we write

1. $|S_j|$
2. Which $h_j : U \to [10|S_j|^2]$ from the canonical p.wise indep. family $\mathcal{H}_j \subseteq \{ f : U \to [10|S_j|^2] \}$

3. For each $\ell \in [10|S_j|^2]$