Hashing

- To maintain $S \subseteq U$, $|S|=n \ll |U|$.  
- Search ($x \in S$?), add, delete from $S$.

Example  $U = \text{All people in the world.}$
$S = \text{All residents of NYC.}$

Trivial:
- Maintain an array of size $|U|$. 
- Too much space.

Hash Table

- Maintain an array $T[0], T[1], \ldots, T[n-1]$. 
- Pick a "hash function" $h : U \rightarrow \{0,1, \ldots, n-1\}$. 
- Store $x \in S$ at location $T[h(x)]$.

Example  person $\rightarrow$ (eye-color, height, nationality).
Collisions - All $x \in S$ sit, $h(x) = i$ are stored at location $T[i]$ in a list.

- Search($x$) takes time $O(k)$ if this list has size $k$.

Search: Given $x$, search list at $T[h(x)]$.

Add: "add to",

Delete: "delete from",

We desire that - Very few collisions
- Sizes of lists are small.

Note: Randomization is necessary, i.e. we cannot pick the hash function $h: U \to \{0, 1, \ldots, n-1\}$ in fixed, a priori, deterministic manner.

Because

$S$ could, adversarially, be such that

$\forall x \in S$, $h(x) = 0$. 
Hence - Randomization!

- Let \( \mathcal{H} \) be a family of functions from \( U \) to \( \{0, 1, \ldots, n-1\} \).

Pick \( h \in \mathcal{H} \) (uniformly) at random.

- Show that for any \( S \subseteq U \), \( |S| = n \), over the choice of \( h \in \mathcal{H} \), few collisions, small lists.

Tradeoff: If \( h \) is truly completely random,

(i.e. \( \mathcal{H} \) is family of all functions \( U \to \{0, 1, \ldots, n-1\} \)) then the scheme "works". However, then "storing" \( h \) takes space proportional to \( 10^1 \).

Hence we desire that

1. \( h \in \mathcal{H} \) is "random enough"
2. \( h \in \mathcal{H} \) has compact representation ("formula")

\[ |\mathcal{H}| \] is "small".
Def A family of functions $H: U \rightarrow \{0,1,\ldots,n-1\}$ is called $2$-universal pairwise independent if

1. $\forall x \in U, \forall i \in \{0,1,\ldots,n-1\}$
   $$\Pr_{h \in H} \left[ h(x) = i \right] = \frac{1}{n}.$$

2. $\forall x, y \in U, x \neq y, \forall i, j \in \{0,1,\ldots,n-1\}$
   $$\Pr_{h \in H} \left[ h(x) = i \land h(y) = j \right] = \frac{1}{n^2}.$$

Note
- $\Rightarrow 1$
- $\Rightarrow \forall x \neq y \in U, x + y,$
  $$\Pr_{h \in H} \left[ h(x) = h(y) \right] = \frac{1}{n}.$$

Theorem There is an explicit, concrete, $2$-universal family of hash functions $H$ and all $h \in H$ are efficiently represented & computed.
Here onwards let $\mathcal{H}$ be a 2-universal family of hash functions $h : U \rightarrow \{0,1,\ldots,n-1\}$.

For $i \in \{0,1,\ldots,n-1\}$, let $L(i)$ denote the list of all elements in $S$ hashed to location $i$.

All probabilities/expectations are over choice of $h \in \mathcal{H}$.

**Lemma**

$$E[|L(i)|] = 1.$$  

**Proof** For every $a \in S$, let $X_a$ be indicator r.v.,

$$X_a = \begin{cases} 1 & \text{if } h(a) = i \\ 0 & \text{otherwise} \end{cases}.$$  

$$E[X_a] = \Pr[h(a) = i] = \frac{1}{n}.$$  

$$|L(i)| = \sum_{a \in S} X_a.$$  

$$\therefore E[|L(i)|] = \sum_{a \in S} E[X_a] = n \cdot \frac{1}{n} = 1.$$
Markov's Inequality

Let \( X \) be a non-negative random variable and \( t \geq 1 \). Then

\[
\Pr [ X \geq t \cdot \mathbb{E}[X] ] \leq \frac{1}{t}.
\]

Lemma

\[
\Pr [ |L(i)| \geq t ] \leq \frac{1}{t}. \quad \text{(Think of } t = 50)\]

Proof.

\[
\mathbb{E}[|L(i)|] = 1.
\]

Markov's inequality.

Chebychev's Inequality

Definition. Let \( X \) be a r.v. Its variance

\[
\text{var}(X) = \mathbb{E}[|X - \mathbb{E}[X]|^2] = \mathbb{E}[|X - \mu|^2] = \mathbb{E}[X^2] - \mu^2.
\]

Fact

\[
\text{var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathbb{E}[X^2] - \mu^2.
\]
Proof \[ \text{var}(x) = E[(x - \mu)^2] \]
\[ = E(x^2 - 2\mu x + \mu^2) \]
\[ = E(x^2) - 2\mu \cdot E(x) + \mu^2 \]
\[ = E(x^2) - \mu^2. \]

Chebychev's Inequality

Let \( X \) be a non-negative r.v. Then

\[ P_X[|X - \mu| \geq T] \leq \frac{\text{var}(x)}{T^2}. \]

Proof \[ P_X[|X - \mu| \geq T] = P_X[|X - \mu|^2 \geq T^2] \]
\[ \leq \frac{E[(X - \mu)^2]}{T^2} \quad \text{Markov} \]
\[ = \frac{\text{var}(x)}{T^2}. \]

Corollary. If \( X \) is a non-negative r.v. Then \( \forall t \geq 1 \)

\[ P_X[ X \geq t \cdot E[X] ] \leq \frac{\text{var}(x)}{(t-1)^2 \mu^2}. \]
Proof

\[ P_r \left[ X \geq t \cdot E[X] \right] \leq P_r \left[ |X - E[X]| \geq (t-1) \cdot E[X] \right] \]

\[ \leq \frac{\text{var}(X)}{(t-1)^2 \cdot E[X]^2} \]

Recall

- \( h \in H \) from 2-universal family,
- \( L(i) = \{ x \in S \mid h(x) = i \} \).
- \( \forall x, y \in U, x+y, \forall i, j \in \{0, \ldots, n-1\} \)
  \[ P_\theta \left[ h(x) = i \land h(y) = j \right] = \frac{1}{n^2} \]
- \( E[|L(i)|] = 1 \).

Claim \( E[|L(i)|^2] \leq 2 \).

Hence \( \text{var}(|L(i)|) = E[|L(i)|^2] - E[|L(i)|]^2 \leq 1 \).
**Proof** \( \forall a \in S \), let \( X_a \) be a r.v.

\[
X_a = \begin{cases} 
1 & \text{if } h(a) = i \\
0 & \text{otherwise}
\end{cases}
\]

\[
\therefore |L(i)| = \sum_{a \in S} X_a.
\]

\[
\therefore \mathbb{E}[|L(i)|^2] = \mathbb{E}\left[\left(\sum_{a \in S} X_a\right)^2\right]
\]

\[
= \mathbb{E}\left[\sum_{a, b \in S} X_aX_b\right]
\]

\[
= \sum_{a \in S} \mathbb{E}[X_a^2] + \sum_{a \neq b} \mathbb{E}[X_aX_b]
\]

\[
= \sum_{a \in S} \Pr[X_a = 1] + \sum_{a \neq b} \Pr[X_a = 1 \land X_b = 1]
\]

\[
= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n^2}
\]

\[
\therefore 2 - \text{uniformity}
\]

\[
\leq 2.
\]
Lemma \quad \Pr \left[ |L(i)| \geq t \right] \leq \frac{1}{(t-1)^2}

Proof. Applying the corollary,

\[
\Pr \left[ |L(i)| \geq t \right] = \Pr \left[ |L(i)| \geq t \cdot \mathbb{E}[|L(i)|] \right] 
\leq \frac{\text{var}(|L(i)|)}{(t-1)^2 \cdot \mathbb{E}[|L(i)|]^2} = \frac{\mu}{(t-1)^2}.
\]

\[\therefore \text{ For every } i, \text{ probability that } |L(i)| \geq 50 \text{ is } \leq \frac{1}{2000}.\]

Note \quad \mathbb{E} \left[ \sum_{i=0}^{n-1} |L(i)|^2 \right] \leq 2n.

- Interpretation: Sum over a.e.s, cost of \text{SEARCH}(a).
- \therefore After hashing, average cost of \text{search}(a) is \text{O}(1).
Example of 2-Universal Hash Family

- Suppose \(|S| = 1\) \(\cup 1 = p\). (prime).

- Consider family of hash functions

\(h_{a,b}: U \rightarrow \{0,1,\ldots, p-1\}\), \(U = \{0,1,\ldots,p-1\}\)

- \(H = \{ h_{a,b} \mid a, b \in \{0,1,\ldots, p-1\}\}\) where

\[ h_{a,b}(x) = ax + b \pmod{p}. \]

- \(|H| = p^2\).

- 2-universality

Fix \(x, y \in U = \{0,1,\ldots, p-1\}\)

\(i, j \in \{0,1,\ldots, p-1\}\).

Then \(h_{a,b}(x) = i \Rightarrow ax + b = i\)

\(h_{a,b}(y) = j \Rightarrow ay + b = j\)

\([\begin{bmatrix} x \\ y \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}] = \begin{bmatrix} i \\ j \end{bmatrix}\)

has unique solution \((a^*, b^*)\).
\[ \Pr \left[ h_{a,b}(x) = i \land h_{a,b}(y) = j \right] = \frac{1}{p^2}. \]

- **Generalization**

- Let \( U = \{0, 1, \ldots, p-1\}^k \).

- \( \mathcal{H} = \{ h_{a_1, \ldots, a_k, b_1, \ldots, b_k} \mid a_1, \ldots, a_k, b_1, \ldots, b_k \in \{0, \ldots, p-1\} \} \).

where

\[ h_{a_1, \ldots, a_k, b_1, \ldots, b_k}(x = (x_1, \ldots, x_k)) = \sum_{i=1}^{k} a_i x_i + b_i \pmod{p}, \]