Hash Tables

Subset of Universe $U = \{0, \ldots, 2^w - 1\} \rightarrow$ Table $T[0 \ldots m-1]$

$$h(x) = \lfloor kx \rfloor \mod m$$

hash function $h: U \rightarrow [m]$ Store $x \in U$ at $T[h(x)]$

This function must be random.

Fiction: $h$ maps $U$ to $[m]$ completely randomly.

Really: Fix in advance a set $H$ of hash functions

- easy to evaluate / Fast
- don't need lots of storage

At runtime, choose $h \in H$ at random for each table.

Statistical properties:

**Uniform:** For all $x \in U$, for all $i \in [m]$

$$\Pr[h(x) = i] = \frac{1}{m}$$

$$h_1(x) = \frac{1}{m} \quad \text{for all } x$$

$$h_2(x) = \frac{2}{m} \quad \text{for all } x$$

$$h_3(x) = \frac{3}{m} \quad \text{for all } x$$

$$H = \{h_1, h_2, \ldots, h_m\}$$

**Universal:** For all $x, y \in U$ where $x \neq y$

$$\Pr[h(x) = h(y)] \leq \frac{1}{m}$$

The time for finding $x$ is $O(1 + E[l(x)])$

where $l(x) = \# \text{ items } y$ s.t. $h(x) = h(y)$

$$E[l(x)] \leq \frac{2n}{m} \quad \text{if } H \text{ is universal}$$
1960s/1970s Carter Wegman

1. multiplicative: choose prime $p > |U|$
   \[ (p) = \mathbb{Z}_p, \quad [p]^+ = \mathbb{Z}_{p-1} \]
   Choose a salt $a \in [p]^+$ at random
   \[ h_a(x) = (a \times x \mod p) \mod m \]
   Near-universal: $\Pr_{a \in [p]^+} [h(x) = h(y)] \leq \frac{2}{m}$

2. multiply-add
   Salt $a \in [p]^+$ and $b \in [p]$
   \[ h_{a,b}(x) = ((a \times x + b) \mod p) \mod m \]
   universal uniform 2-uniform

3. binary multiplication
   Salt: $a \in [2^w]$, $|U| = 2^w$, $m = 2^l$
   \[ h_a(x) = \left\lfloor \frac{(a \times x) \mod 2^w}{2^{w-l}} \right\rfloor \]
   near-universal + $b$: 2-uniform

\[(a) \times (x) \Rightarrow \text{(WORDSIZE - HASHBITS)}\]
**Tabulation hashing** \(|U| = 2^w = 2^{w/2} \cdot 2^{w/2} \quad m = 2^l\)

Define two arrays \(A[0..2^{w/2}-1] \quad B[0..2^{w/2}-1]\)
Filled with random values in \([m]\)

\[
h_{A,B}(x,y) = A[x] \oplus B[y]
\]

**Universal:** Fix \((x,y) \neq (x',y')\) wlog \(x \neq x'\)

\[
h(x,y) = h(x',y')
\]

\[
A[x] \oplus B[y] = A[x'] \oplus B[y']
\]

Fix \(A\) and \(B\) except \(A[x]\)

\[
\Pr\left[A[x] = A[x'] \oplus B[y] \oplus B[y']\right] = \frac{1}{m}
\]

Not 4-uniform

Consider items \((x,y), (x,y'), (x',y), (x',y')\)

\[
h(x,y) \oplus h(x',y) \oplus h(x',y') = h(x',y')
\]

For any \(x\), \(E(l(x)) = O(1)\) so \(E[\text{time to find } x] = O(1)\)

Want \(E[\max_x \text{Time}(x)] = O(1)\)

Even if we use ideal random hashing

\[
\max_x l(x) = \Theta\left(\frac{\log n}{\log \log n}\right) \text{ with high prob. assuming } m=n
\]

If \(m=n^2\)

\[
E[\#\text{collisions}] \leq \frac{1}{m} \binom{n^2}{2} < \frac{1}{2} \quad (\text{universal})
\]

\[
\Pr[\text{any collisions}] < \frac{1}{2}
\]
"Perfect hashing"

\[ m = n \]

\[ n_i = \#\{x : h(x) = i\} \]

\[ m_i = n_i^2 \]

WORST case time for search is \( O(1) \)

Assuming primary hash function is universal

\[ E\left[ \sum_i n_i^2 \right] \leq 2n \]

\[
E\left( \sum_i n_i^2 \right) = \sum_i \left( \sum_x [h(x) = i] \right)^2 \\
= \sum_i \left( \sum_x \sum_y [h(x) = i] \cdot [h(y) = i] \right) \\
= \sum_i \left( \sum_x [h(x) = i] + 2 \sum_{x \neq y} [h(x) = h(y) = i] \right) \\
= n + 2 \sum_{x \neq y} [h(x) = h(y)] = n + 2 \left( \frac{1}{n} \right) \leq 2n \]