COL 351: Analysis and Design of Algorithms

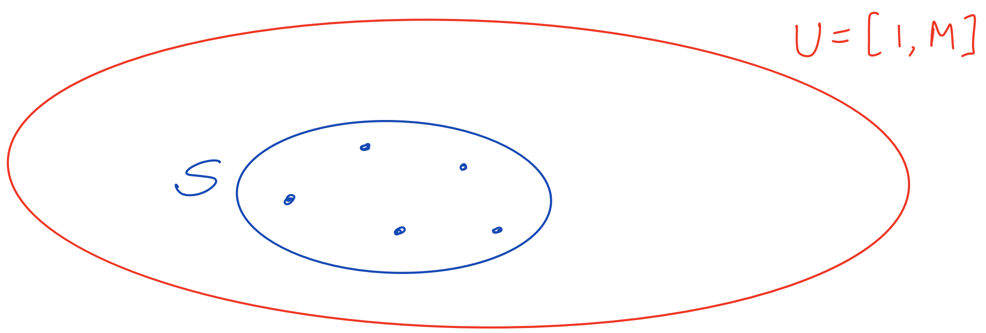
Lecture 20

Set Membership

Given: A universe $U = [1, 2, \ldots, M]$, and a set $S \subsetneq [1, M]$ of size n.

Goal: Find a data-structure of $O(n = |S|)$ size that answers for any $x \in [1, M]$ query of form:

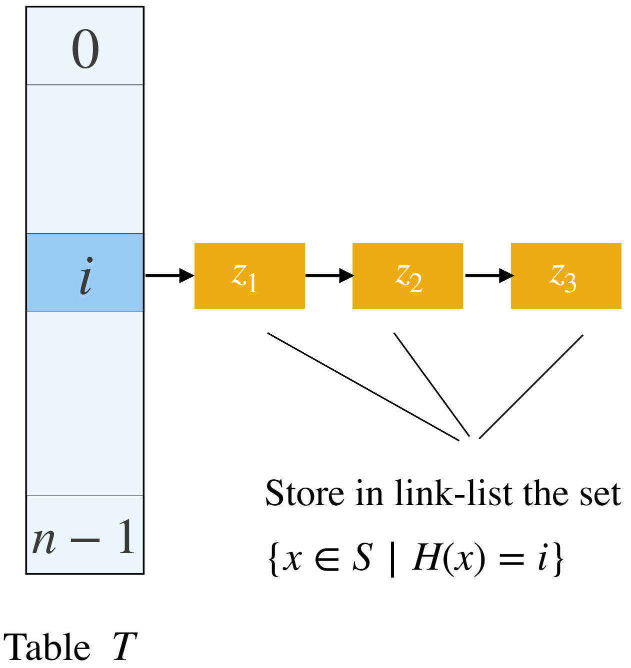
"Does $x \in S$?"



Hash Function

$H(z) = z \mod n$

- Works well for a random S
- What if S is <u>not</u> random?



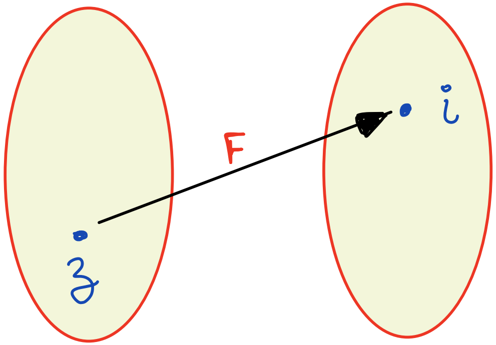
Claim: No single hash function can work for all possible sets S

Modular Arithmetic

$F(z) \; = \; (r \cdot z) \; \bmod p$ (Here, p is a prime).

Claim: If $r \in [1, p-1]$ was random, then for any $z, i \in [1, p-1]$, we have

$$\mathrm{Prob}\big(F(z) = i\big) = \frac{1}{p-1}.$$

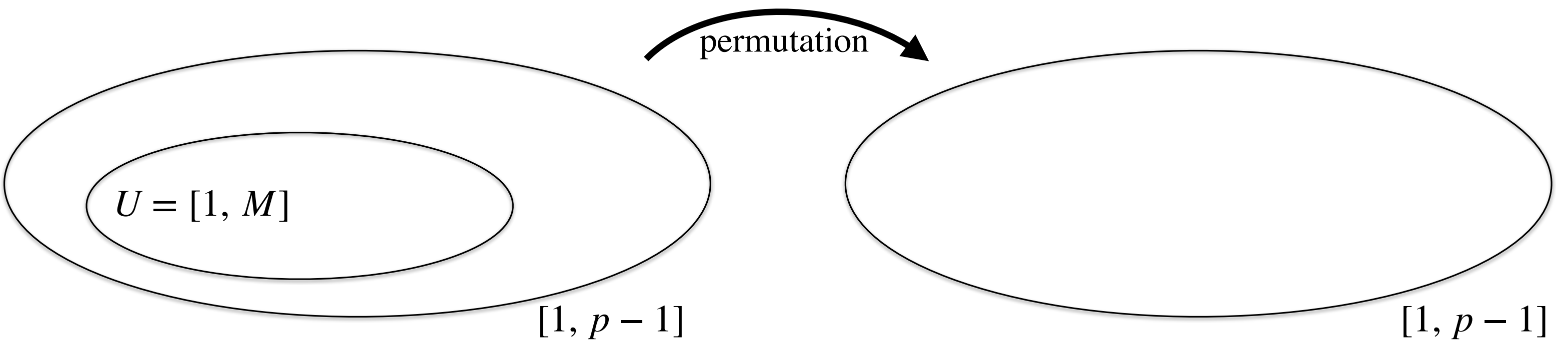


New Hash Function

- Universe $U = [1, M]$.
- p = prime in range $[M+1, 2M]$, r = integer range $[1, p-1]$

Hash Function:

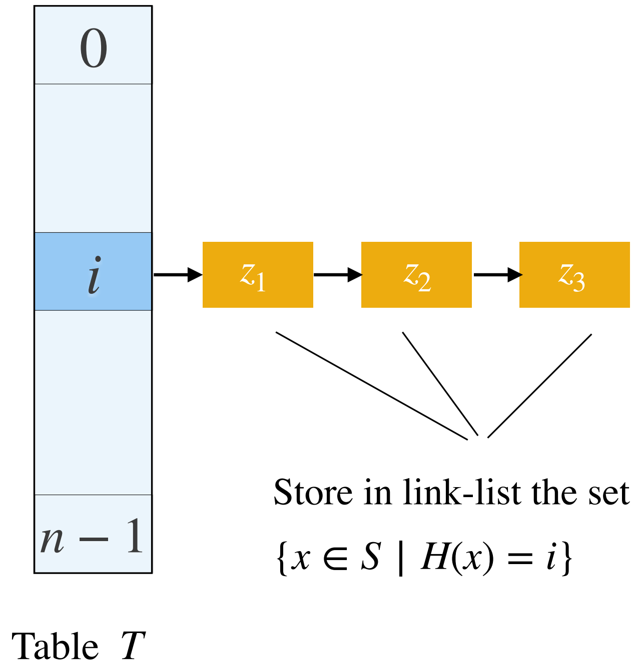
$$H_r(z) = \big((r \cdot z) \mod p\big) \mod n$$



New Hash Function

Hash Function:

$$H_r(z) \;=\; \big((r \cdot z) \mod p\big) \mod n$$



What is collision probability?

Hash Function:

$$H_r(z) \;=\; \big((r \cdot z) \mod p\big) \mod n$$

Question: For any $x, y \in U$, what is **collision probability** if 'r' is randomly chosen?

Solution:

$$H_r(x) = H_r(y)$$

$$\Rightarrow\ (rx \mod p) - (ry \mod p) \text{ is multiple of } n$$

$$\Rightarrow\ (rx - ry \mod p) \text{ is multiple of } n \quad \textbf{or} \quad (rx - ry \mod p) - p \text{ is multiple of } n$$

$$\Rightarrow\ (rx - ry \mod p) \text{ lies in } \{n,\ 2n,\ 3n,\ \ldots\} \quad \textbf{or} \quad \{p-n,\ p-2n,\ p-3n,\ \ldots\}$$

$$\Rightarrow\ (r\ (x - y) \mod p) \text{ lies in } \{n,\ 2n,\ 3n,\ \ldots,\ p-3n,\ p-2n,\ p-n\}$$

$$\text{Prob}\Big(H_r(x) = H_r(y)\Big) \ \leq \frac{1}{p-1} \cdot |\{n,\ 2n,\ 3n,\ \ldots,\ p-3n,\ p-2n,\ p-n\}| \ \approx\ \frac{2}{n}$$

Expected Time to search an element

***Question*:** For any $x \in U$, what is expected time to verify membership of x in set S?

Solution:

The time to search x is sum of

(i) Time to compute $H_r(x)$, and

(ii) Number of elements in S mapped to $H_r(x)$.

Expected Time:

$$= 1 + \sum_{y \in S \setminus \{x\}} \text{Prob}\big(H_r(y) = H_r(x)\big) \leq 1 + (n-1) \cdot \frac{2}{n} = O(1)$$

Total number of Collisions

***Question*:** What is expected number of total collisions?

Solution:

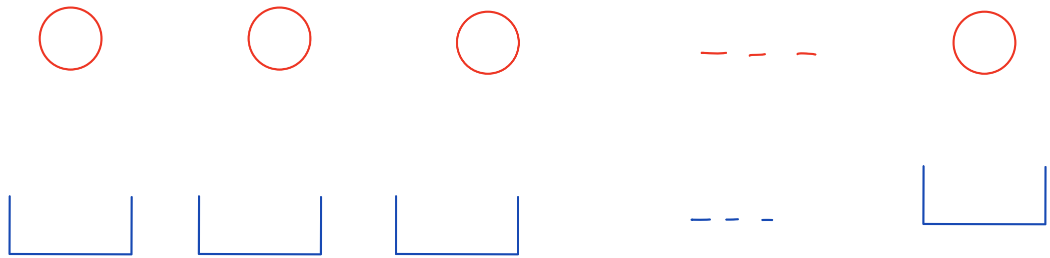
Expected total number of collisions are

$$= \sum_{\substack{x, y \in S \\ x \neq y}} \operatorname{Prob}\big(H_r(y) = H_r(x)\big) \;\leq\; \frac{n(n-1)}{2} \cdot \frac{2}{n} \;\leq\; n$$

Balls - Bins Exercise

n balls

n bins



- each ball goes into one of the randomly selected Bin
- Expected no of balls in Bin i $= \frac{1}{n} \cdot n = 1$
- Fact

$$\text{Exp}\left(\max_{i=1}^{n} \left(\# \text{ of Balls in Bin } i\right)\right) = \Theta\left(\frac{\log n}{\log \log n}\right)$$

sol^n to $k^k = n$

Maximum Time to search an element

Worst-Case time

***Question*:** What is expected value of $\max_{i \in [0,\, n-1]} |T[i]|$?

Is hashing any better than AVL trees?

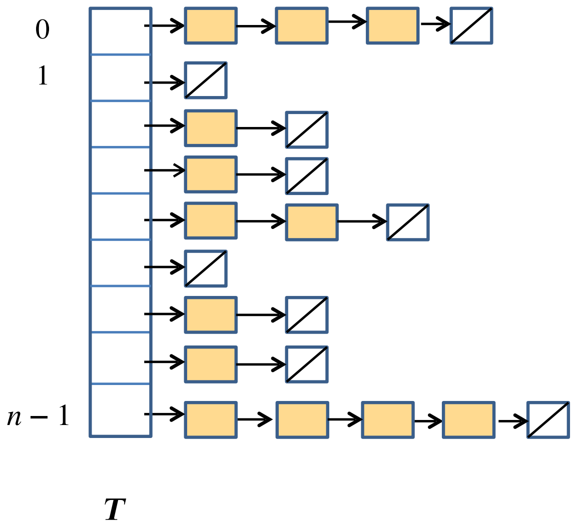
<u>Answer:</u>

It will be $\Theta\left(\frac{\log n}{\log \log n}\right)$.

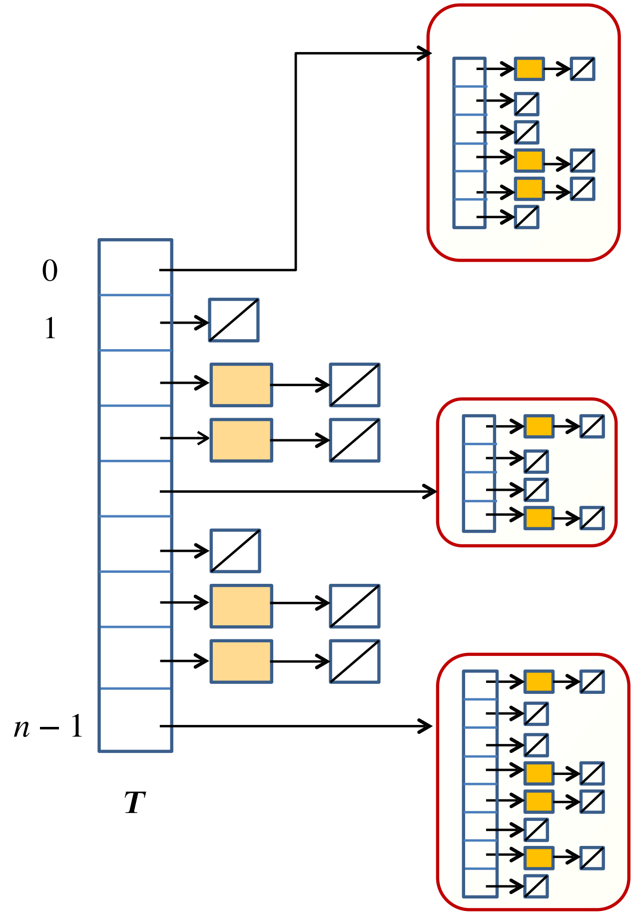
<u>Remark</u>:

The proof of the fact $Exp\left(\max_{0 \le i \le n-1} |T[i]|\right) = \Theta\left(\frac{\log n}{\log \log n}\right)$ is not part of Syllabus.

Two-Level Hash Table



Old Table



New Table

Two-Level Hash Table

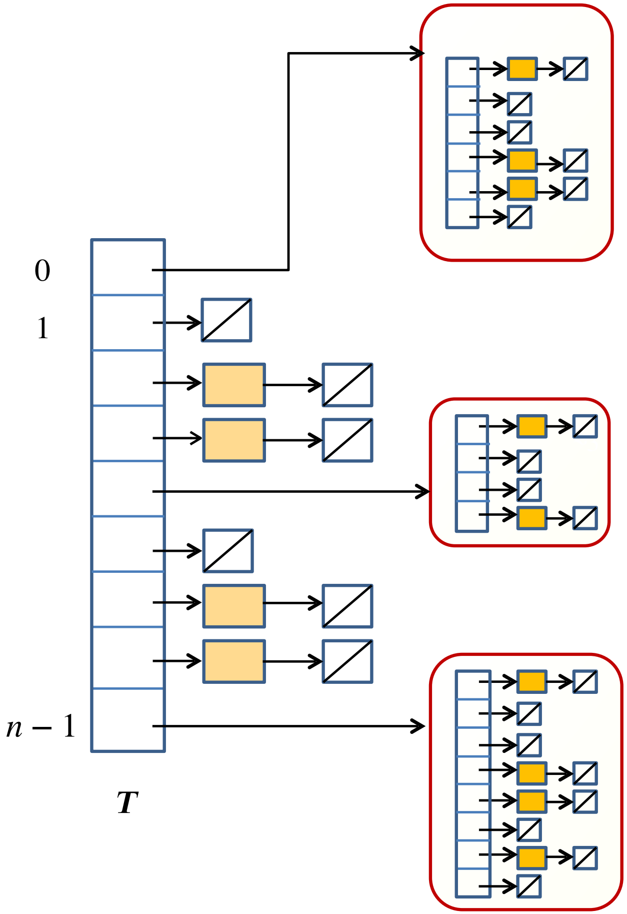
Outer Hash Function:

$$H_r(z) \;=\; \big((r \cdot z) \mod p\big) \mod n$$

Inner Hash Function:

$$z \mapsto \big((r_0 \cdot z) \mod p\big) \mod n_i^2$$

where, $n_i =$ size of $T[i]$



New Table

Two-Level Hash Table

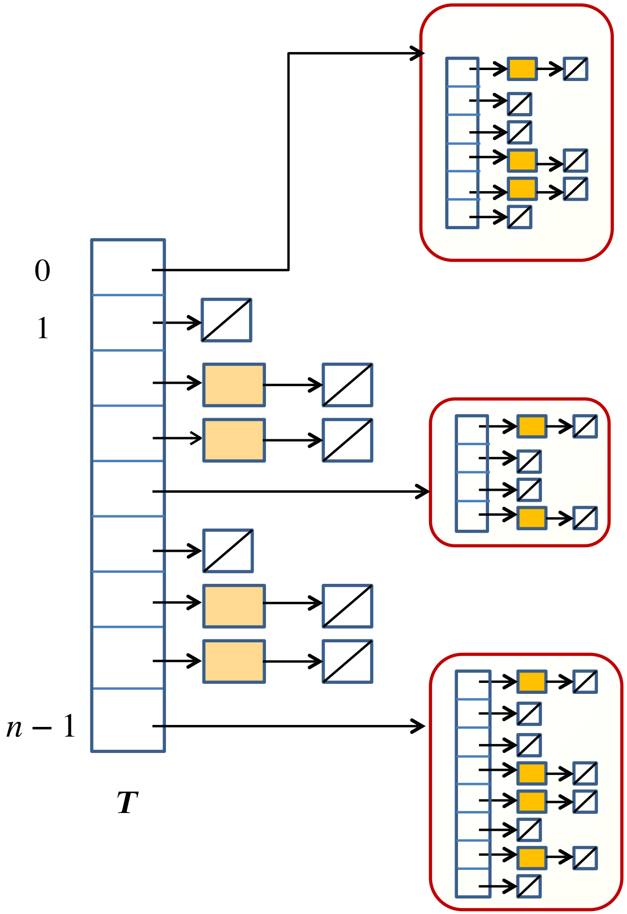
Outer Hash Function:

$H_r(z) \;=\; \big((r \cdot z) \mod p\big) \mod n$

Inner Hash Function:

$z \mapsto \big((r_0 \cdot z) \mod p\big) \mod n_i^2$

where, n_i = size of $T[i]$



New Table

- What is expected total size?
- What is expected number of total collisions?