Setting:

Universe $U$ = set of all possible values

Subset $S$ = interested in this subset

$|S| = N$

w.h.p. elements from $S$ have small number of collisions

$x, y \in S$

$\exists \text{ collision } \Rightarrow h(x) = h(y)$
What is random?

1. inputs
2. hash function

E.g. NTP switch: IP address X.X.X.X

Assume a family of hash functions \( H \)

When time to hash "S", we choose a random function

\( h \in H \)
\[ M = \text{table size} \]
\[ \{M\} = \{0, 1, \ldots, M-1\} \]

What properties?

1. Small probability of distinct keys colliding:
   \[ \text{If } x \neq y, \quad P[h(x) = h(y)] \text{ is "small"} \]

2. \( h \) is easy to compute

3. \( h \) is easy to store (small num of bits)

4. \( M \) is small \( \Rightarrow \) (hash table size is small)
Ideal Hash Function

Perfectly random:
for each \( x \in S \), \( h(x) = \) a uniformly random location in \([M]\)

Properties:
- Low collision prob. \( P[h(x) = h(y)] = \frac{1}{M} \) for any \( x \neq y \)
- Even conditioned on hashed values for any other subset \( A \) of \( S \)

Downsides:
- Too large to store
- Compute: table look up
Universal Hash Functions

Captures the property of non-collision of two distinct elements.

**Defn:** A family \( \{ h \} \) of hash functions mapping \( U \rightarrow [m] \) is universal if for any \( x \neq y \)

\[
P[ h(x) = h(y) ] \leq \frac{1}{M}
\]

needs to hold for every pair \( x \leq y, x \neq y \)
Simple construction of universal hashing: 

Let $|U| = 2^u$ elements are (vectors of length $u$).

Let $|M| = 2^m$.

Let $A$ be binary matrix

$A$ is a uniform random binary matrix.

For any $x \in U$, $h(x) := Ax \mod 2^m$.

Q: How many hash functions?

$2^m$
**Thm.** This family of hash function is universal.

**Proof:**

\[ h(x) = h(y) \quad \text{where} \ x \neq y \]

\[ A \cdot x = A \cdot y \]

\[ A \cdot (x - y) = 0 \]

\[ A \cdot z = 0 \quad \text{for} \ z \neq 0 \]

\[ \text{To show:} \quad P(Az = 0) \leq \frac{1}{m} \quad \text{for any} \ z \neq 0 \]

\[ Az = 0 \quad \Rightarrow \quad \sum A \cdot z_j = 0 \]

\[ m \times n \]

\[ \begin{bmatrix} A_1 & A_2 & \cdots & A_n \end{bmatrix} \]

\[ \text{Length in} \]

\[ \begin{bmatrix} z_1 \end{bmatrix} \]

\[ \begin{bmatrix} z_2 \end{bmatrix} \]

\[ \begin{bmatrix} z_k \end{bmatrix} \]
Let \( z_{i*} \neq 0 \) (there exists \( i^* \) : \( z_{i^*} = z \neq 0 \))

\[ A_{z} = 0 \quad A_{z} = 0 \]

\[ \Rightarrow \quad A_{i*} = -\sum_{j \neq i^*} A_{j} z_{j} \]

\[ p(A_{z} = 0) = p (A_{i^*} = -\sum_{j \neq i^*} A_{j} z_{j}) \]

\[ = \left( \frac{1}{2} \right)^m \]

\[ = \frac{1}{M} \]

\[ \sum_{j \neq i^*} A_{j} z_{j} \]

\[ A_{i^*} z_{i^*} = -\sum_{j \neq i^*} A_{j} z_{j} \]
Application 1: Hash tables

Closed addressing

Open addressing

Look up time: number of collisions.

Let \( C_x \) = number of other elements mapped to the value where \( x \) is mapped to.

\[
L_x = C_x + 1
\]

\[
E[L_x] = E[C_x] + 1 = \frac{(N-1)}{M} + 1
\]

\[
E[L_x] \leq 2
\]

If \( M \geq N \)
Let $C =$ total number of collisions.

$$E[C] \leq \binom{N}{2} \cdot \frac{1}{M} \cdot \frac{N(N-1)}{2}$$

Collision-free hash table?

If $M \geq N^2$

$$P\left[\text{there exists a collision}\right] = \frac{1}{2}$$

Repeat the expr to get collision-free \( \Rightarrow \) Const. look-up time in worst-case

Downside: $M \geq N^2$

Q. Can we get collision-free with $M = O(N)$?