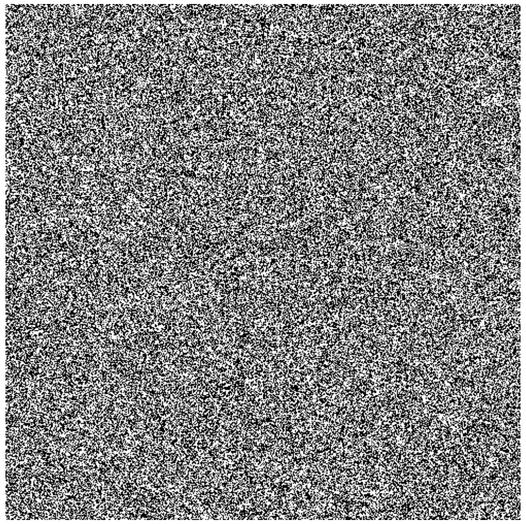
Universal hashing

Ben Langmead

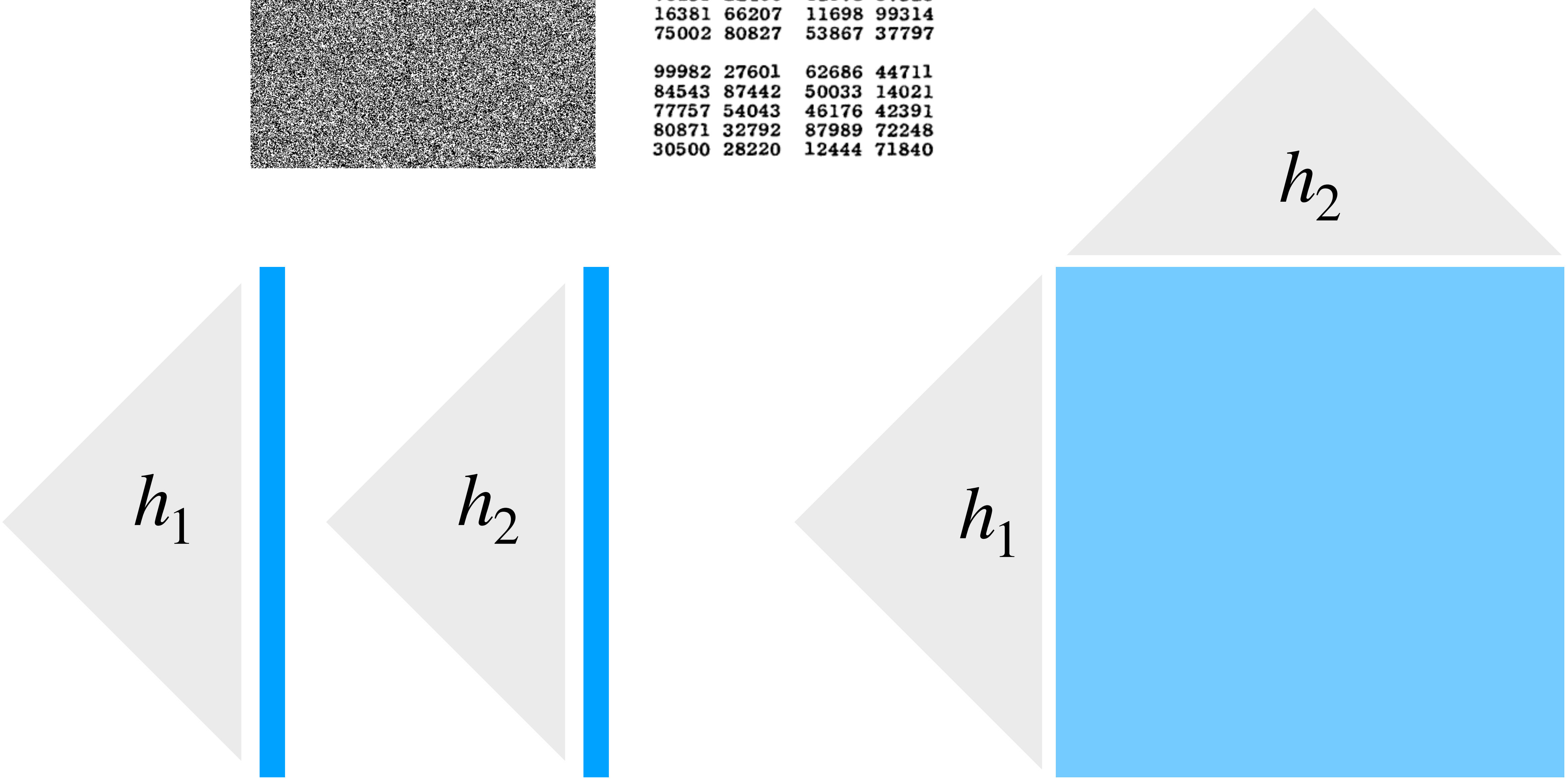
JOHNS HOPKINS
WHITING SCHOOL *of* ENGINEERING

Department of Computer Science

Randomness & independence



73735	45963	78134	63873
02965	58303	90708	20025
98859	23851	27965	62394
33666	62570	64775	78428
81666	26440	20422	05720
15838	47174	76866	14330
89793	34378	08730	56522
78155	22466	81978	57323
16381	66207	11698	99314
75002	80827	53867	37797
99982	27601	62686	44711
84543	87442	50033	14021
77757	54043	46176	42391
80871	32792	87989	72248
30500	28220	12444	71840

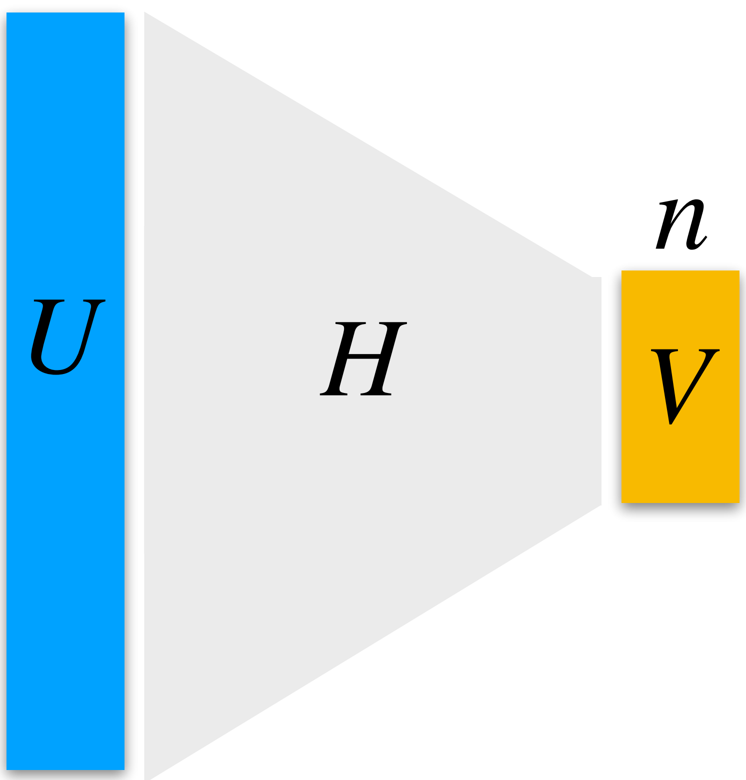


Universal hashing

A family of hash functions H from universe U with $|U| \geq n$ to range {0, 1, ... , $n-1$} is **2-universal** if

for distinct elements x_1, x_2 and for function h drawn uniformly from H:

$$\Pr\left(h(x_1) = h(x_2)\right) \leq \frac{1}{n}$$



Let's prove a useful expectation for hash tables…

Universal hashing

A set S of m items have been hashed to an n-bucket hash table using h from a 2-universal family

For given element x let r.v. X be the number of items in bucket $h(x)$. We want to show:

$$\mathbf{E}[X] \leq \begin{cases} m/n & \text{if } x \notin S \\ 1 + (m-1)/n & \text{if } x \in S \end{cases}$$

Not-in-table case

1 if $m = n$

In-table case

< 2 if $m = n$

Universal hashing

$$\mathbf{E}[X] \leq \begin{cases} m/n & \text{if } x \notin S \\ 1 + (m-1)/n & \text{if } x \in S \end{cases}$$

Let X_i be a r.v. $X_i = 1$ when the i^{th} element of S is in same bucket as x. $X_i = 0$ otherwise

$$\Pr(X_i = 1) \leq \frac{1}{n}$$ By 2-universality!

Universal hashing

$x \notin S$ case

Linearity

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{m} X_i\right] = \sum_{i=1}^{m} \mathbf{E}[X_i] \leq \frac{m}{n}$$

2-universality

+ expectation of indicator

$$\mathbf{E}[X_i] = \Pr(X_i = 1) \leq \frac{1}{n}$$

Universal hashing

$$\mathbf{E}[X] \leq \begin{cases} m/n & \text{if } x \notin S \\ 1 + (m-1)/n & \text{if } x \in S \end{cases}$$

Let X_i be a r.v. $X_i = 1$ when the i^{th} element of S is in same bucket as x. $X_i = 0$ otherwise

Without loss of generality, use $i = 1$ for item x

$$\Pr(X_i = 1) \leq \frac{1}{n} \quad \text{for } i > 1$$

Universal hashing

$x \in S$ case

Linearity

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{m} X_i\right] = 1 + \sum_{i=2}^{m} \mathbf{E}[X_i] \leq 1 + \frac{m-1}{n}$$

2-universality

+ expectation of indicator

Universal hashing

Proving a key property; with 2-universal hashing, expected query time is $O(1)$ when $m \le n$

$$\mathbf{E}[X] \le \begin{cases} m/n & \text{if } x \notin S \\ 1 + (m-1)/n & \text{if } x \in S \end{cases}$$

Not-in-table case

1 if $m = n$

In-table case

~2 if $m = n$

Universal hashing

What kind of family has this property?

Are functions easy to draw from the family?

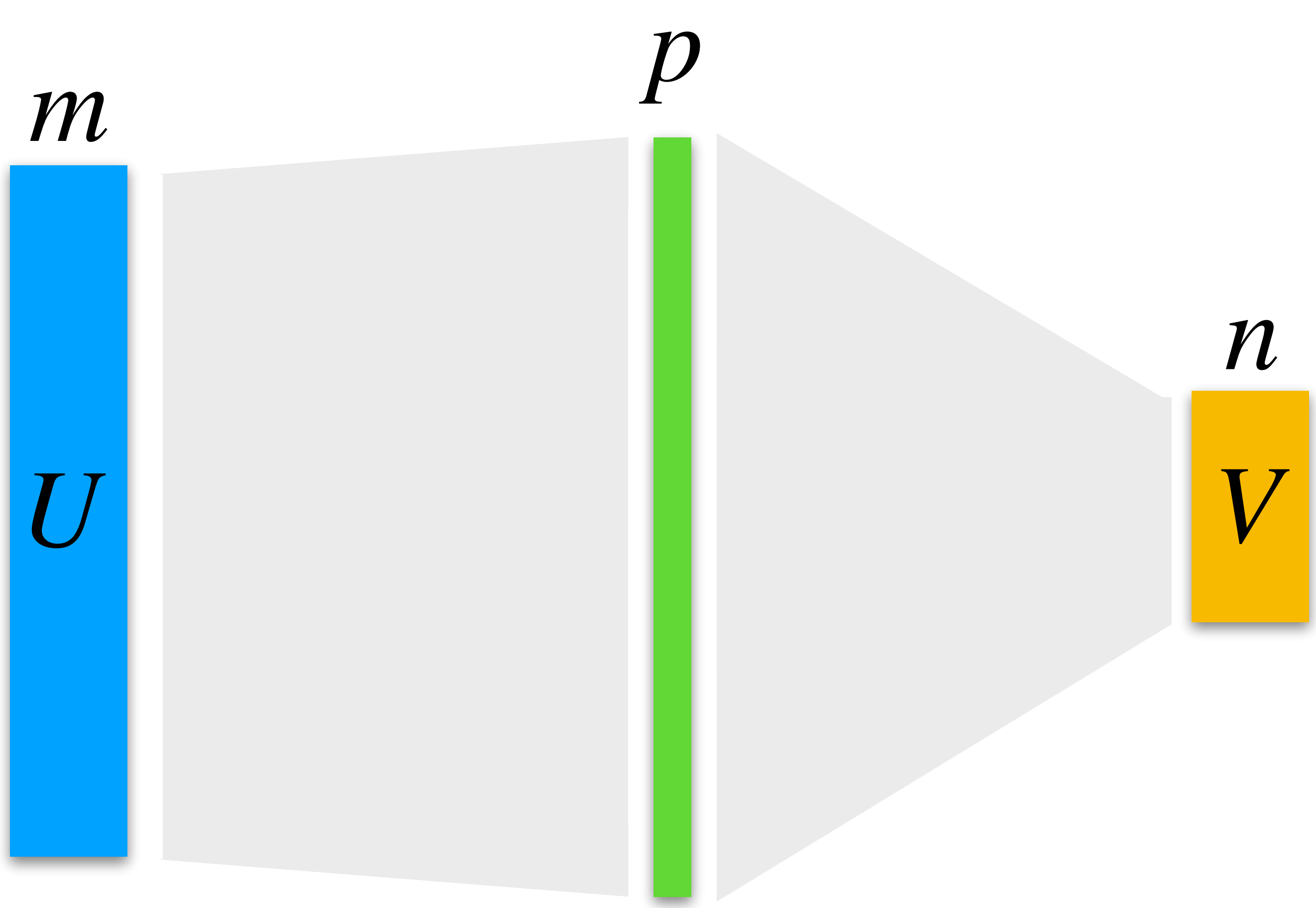
Are functions easy to store and compute with?

Universal hashing

Universe $U : \{0, 1, 2, \ldots, m-1\}$

Range $V : \{0, 1, 2, \ldots, n-1\}$ with $n \leq m$

Prime $p \geq m$

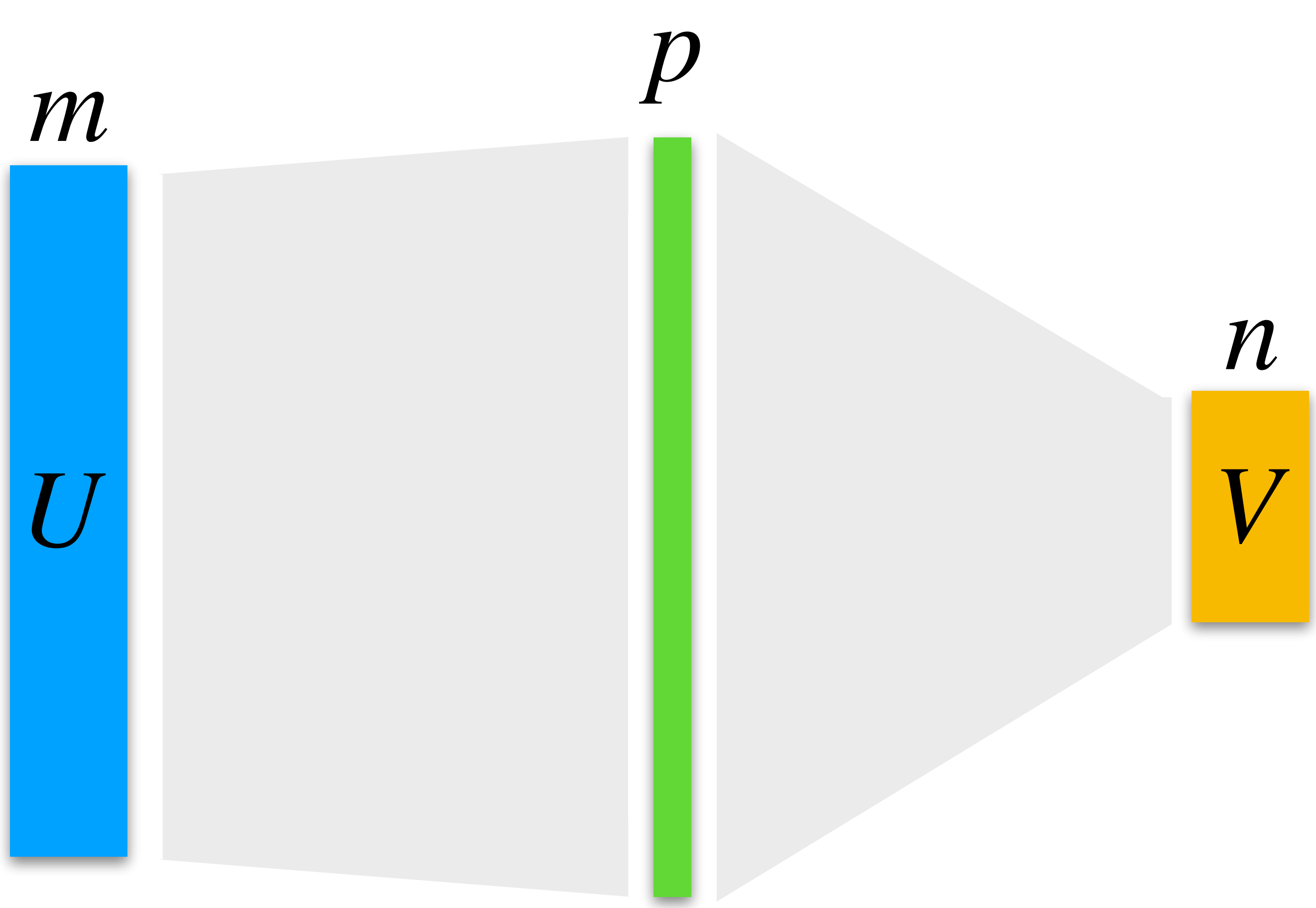


Universal hashing

Example of a 2-universal family from U to V:

$$H = \{h_{a,b} \mid \boxed{1} \leq a \leq p - 1,\ 0 \leq b \leq p - 1\}$$

$$h_{a,b}(x) = ((ax + b) \mod p) \mod n$$



Prime field

A prime field $\mathbf{F}_p$ is a number system consisting of integers modulo a prime p, and rules for plus & times

Plus & times have many of our favorite properties

$\mathbf{F}_5$

+	0	1	2	3	4
0	0	1	2	3	4
1	1	2	3	4	0
2	2	3	4	0	1
3	3	4	0	1	2
4	4	0	1	2	3

x	0	1	2	3	4
0	0	0	0	0	0
1	0	1	2	3	4
2	0	2	4	1	3
3	0	3	1	4	2
4	0	4	3	2	1

Prime field

Fields are special for having ***multiplicative inverses***

Each number (except 0) has another it multiplies with to get 1

$$2 \cdot 3 = 3 \cdot 2 = 1 \mod 5$$
$$4 \cdot 4 = 1 \mod 5$$
$$1 \cdot 1 = 1 \mod 5$$

$\mathbf{F}_5$

x	0	1	2	3	4
0	0	0	0	0	0
1	0	**1**	2	3	4
2	0	2	4	**1**	3
3	0	3	**1**	4	2
4	0	4	3	2	**1**

Prime field

Does modulo a non-prime work?

$\mathbf{F}?_6$

x	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	**0**	2	4
3	0	3	**0**	3	**0**	3
4	0	4	2	**0**	4	2
5	0	5	4	3	2	1

Signs of trouble. 1) We sometimes get 0s when multiplying non-0s

Prime field

Does modulo a non-prime work?

$\cancel{\mathbf{F}}?_6$

x	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	**0**	2	4
3	0	3	**0**	3	**0**	3
4	0	4	2	**0**	4	2
5	0	5	4	3	2	1

no 1

Signs of trouble. 1) We sometimes get 0s when multiplying non-0s

2) Some rows don't have 1; no multiplicative inverse

Prime field

$\mathbf{F}_7$

x	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

Universal hashing

Choose distinct $x_1, x_2 \in U$. Can they collide in p?

$$ax_1 + b \stackrel{?}{=} ax_2 + b \mod p$$

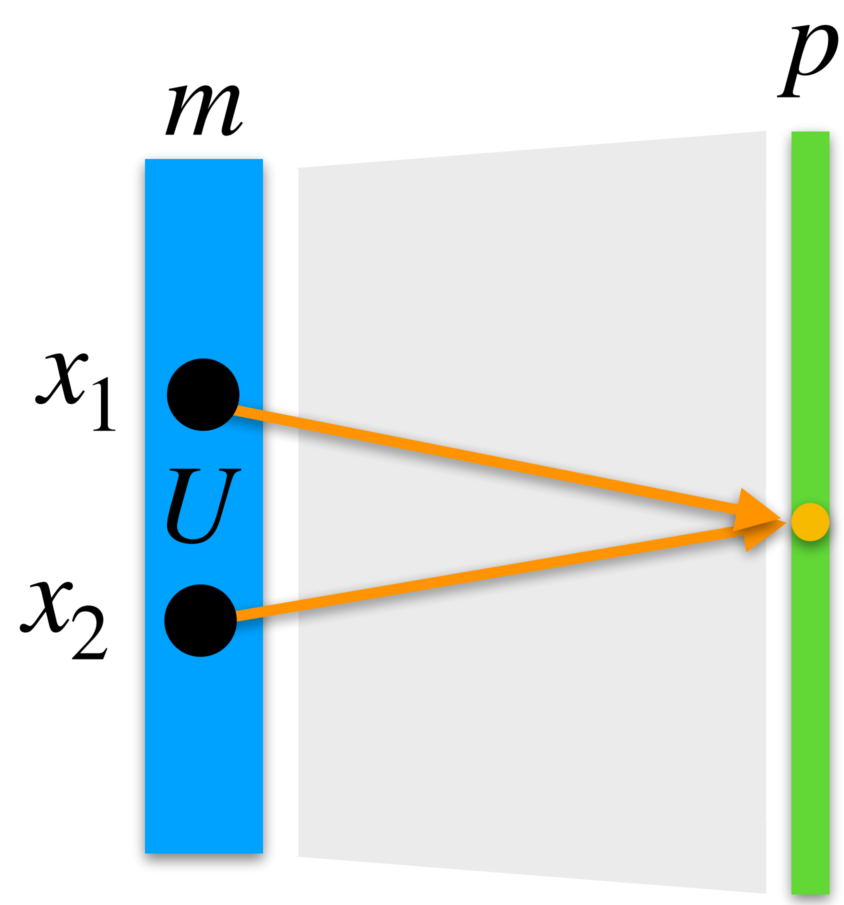
$$ax_1 + b = ax_2 + b \mod p$$

$$ax_1 = ax_2 \mod p$$

$$a(x_1 - x_2) = 0 \mod p$$

We said $a \geq 1$ and $x_1 \neq x_2$

Left side is product of two numbers and neither is 0 mod p.



Prime field

$\mathbf{F}_7$

x	0	1	2	3	4	5	6
0	0	0	0	0	0	0	0
1	0	1	2	3	4	5	6
2	0	2	4	6	1	3	5
3	0	3	6	2	5	1	4
4	0	4	1	5	2	6	3
5	0	5	3	1	6	4	2
6	0	6	5	4	3	2	1

No 0s

Universal hashing

Can $ac = zp$, where p is a prime, z is some integer multiple, and a & c are ***not*** 0 mod p?

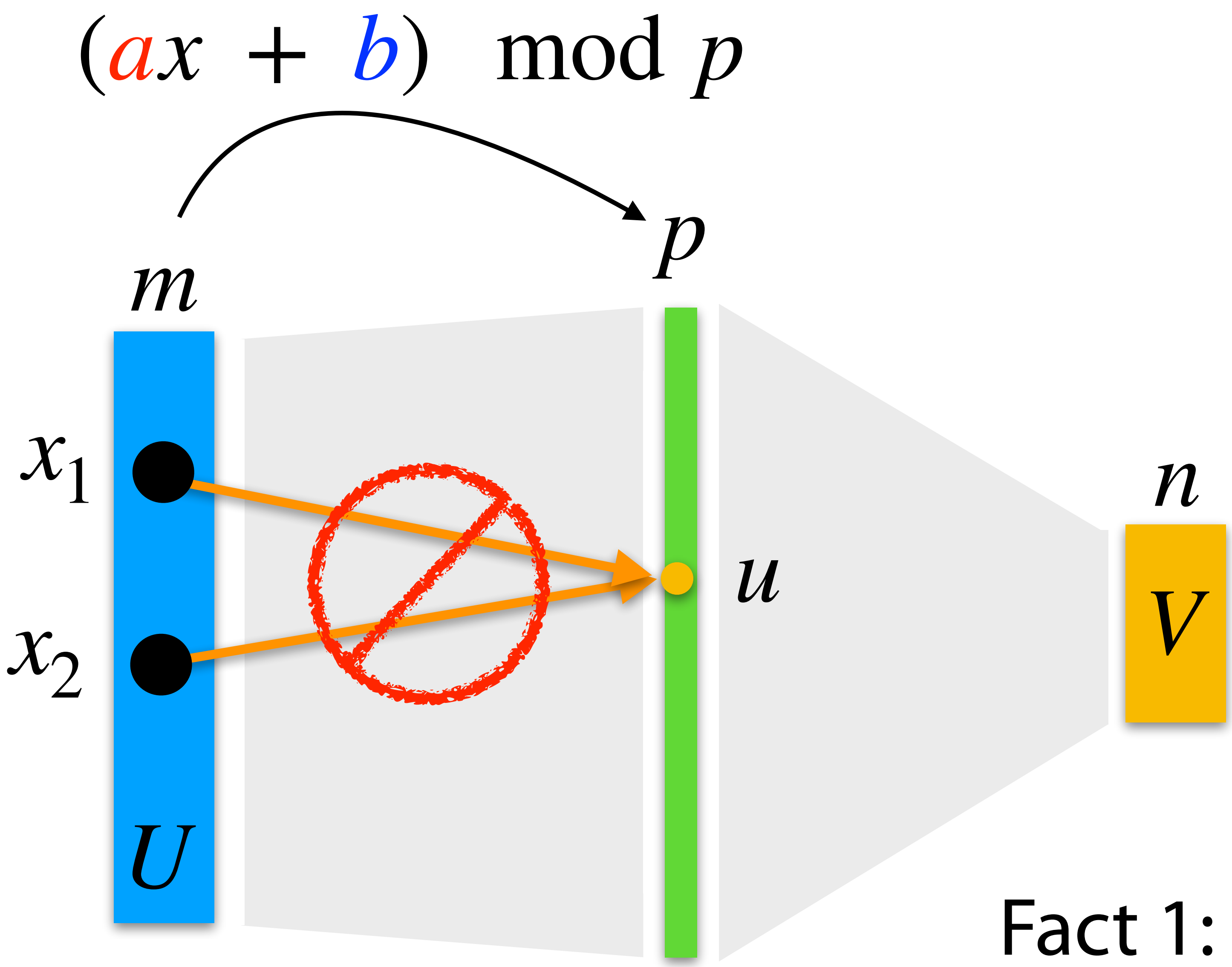
$$ac = zp$$

Consider prime factorizations of a and c

For equality to hold, p must be a prime factor of a or c, contradicting "a & c are ***not*** 0 mod p"

$$ac \neq zp$$

Universal hashing



Fact 1: Distinct items from U
won't collide in prime field

Universal hashing

	b	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
	a	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	0	0	0	0	0																
	1	1	2	3	4																
x	2	2	4	1	3																
	3	3	1	4	2																
	4	4	3	2	1																

Each column is a permutation of integers mod 5

Copied from $\mathbf{F}_5 \times$ table

$$(ax + b) \mod 5$$

Universal hashing

b	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
a	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
x = 0	0	0	0	0	1	1	1	1												
1	1	2	3	4	2	3	4	0												
2	2	4	1	3	3	0	2	4												
3	3	1	4	2	4	2	0	3												
4	4	3	2	1	0	4	3	2												

Same as block to left but + 1 mod 5

$$(ax + b) \mod 5$$

Universal hashing

	b	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
	a	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	0	0	0	0	0	1	1	1	1	2	2	2	2								
	1	1	2	3	4	2	3	4	0	3	4	0	1								
x	2	2	4	1	3	3	0	2	4	4	1	3	0								
	3	3	1	4	2	4	2	0	3	0	3	1	4								
	4	4	3	2	1	0	4	3	2	1	0	4	3								

Same as block to left but + 1 mod 5

$$(ax + b) \mod 5$$

Universal hashing

	b	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
	a	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	0	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
	1	1	2	3	4	2	3	4	0	3	4	0	1	4	0	1	2	0	1	2	3
x	2	2	4	1	3	3	0	2	4	4	1	3	0	0	2	4	1	1	3	0	2
	3	3	1	4	2	4	2	0	3	0	3	1	4	1	4	2	0	2	0	3	1
	4	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0	4	3	2	1	0

$$(ax + b) \mod 5$$

Every column is a permutation of integers mod 5.

Therefore: no collisions. Distinct xs get distinct answers

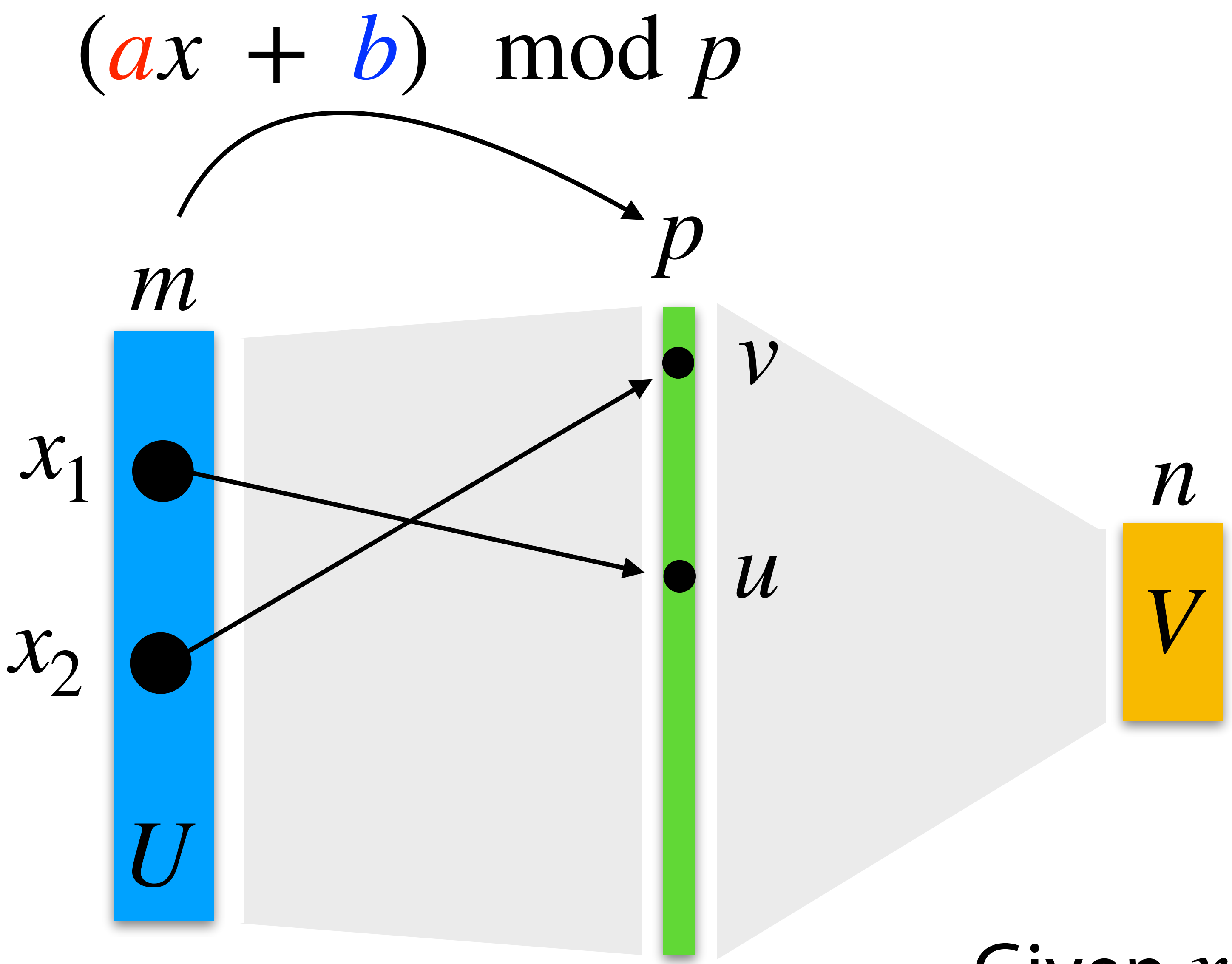
Universal hashing

	b	0	**0**	0	1	1	1	2	2	2	3	3	3
	a	1	**2**	3	1	2	3	1	2	3	1	2	3
x	0	0	**0**	0	1	1	1	2	2	2	3	3	3
	1	1	**2**	3	2	3	0	3	0	1	0	1	2
	2	2	**0**	2	3	1	3	0	2	0	1	3	1
	3	3	**2**	1	0	3	2	1	0	3	2	1	0

$$(ax + b) \mod 4$$

Is every columns necessarily a permutation of another column?

Universal hashing



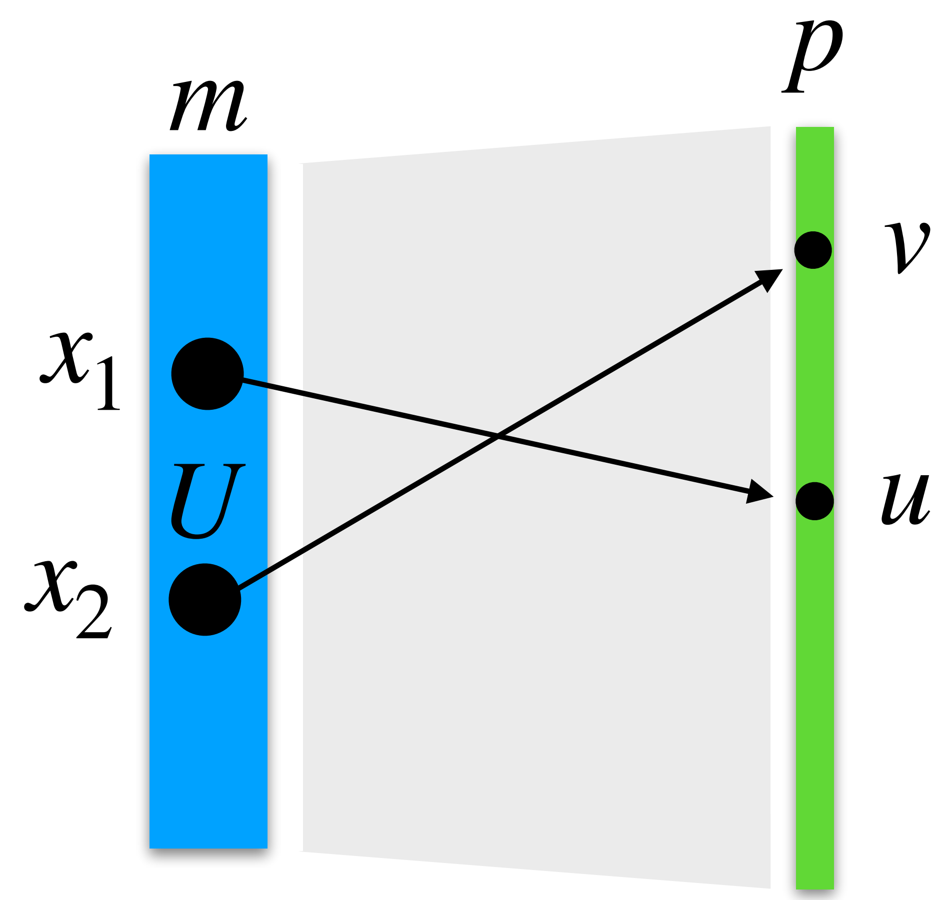
Given x_1, x_2, u, v, what is the chance that $h_{a,b}(x_1) = u$ and $h_{a,b}(x_2) = v$?

Universal hashing

$$(a\,x_1 + b) = u \mod p$$
$$(a\,x_2 + b) = v \mod p$$

$$a = \frac{v-u}{x_2 - x_1} \mod p$$

$$b = u - a x_1 \mod p$$



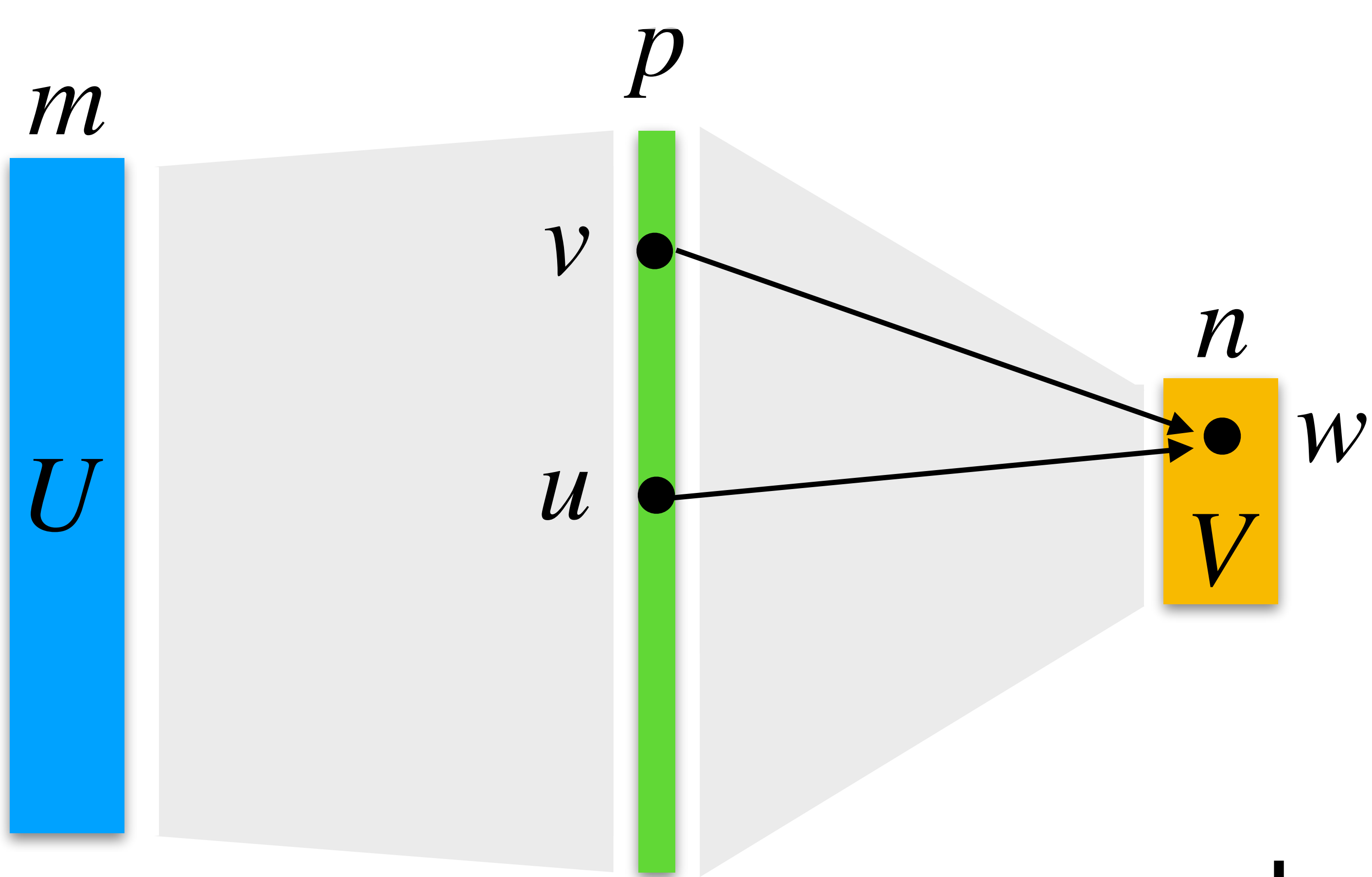
Fact 2: Single choice of a, b satisfies the equations.

$0 \le b \le p-1$ and $1 \le a \le p-1$, so chance is $\dfrac{1}{p(p-1)}$

u, v pairs are equally likely

Universal hashing

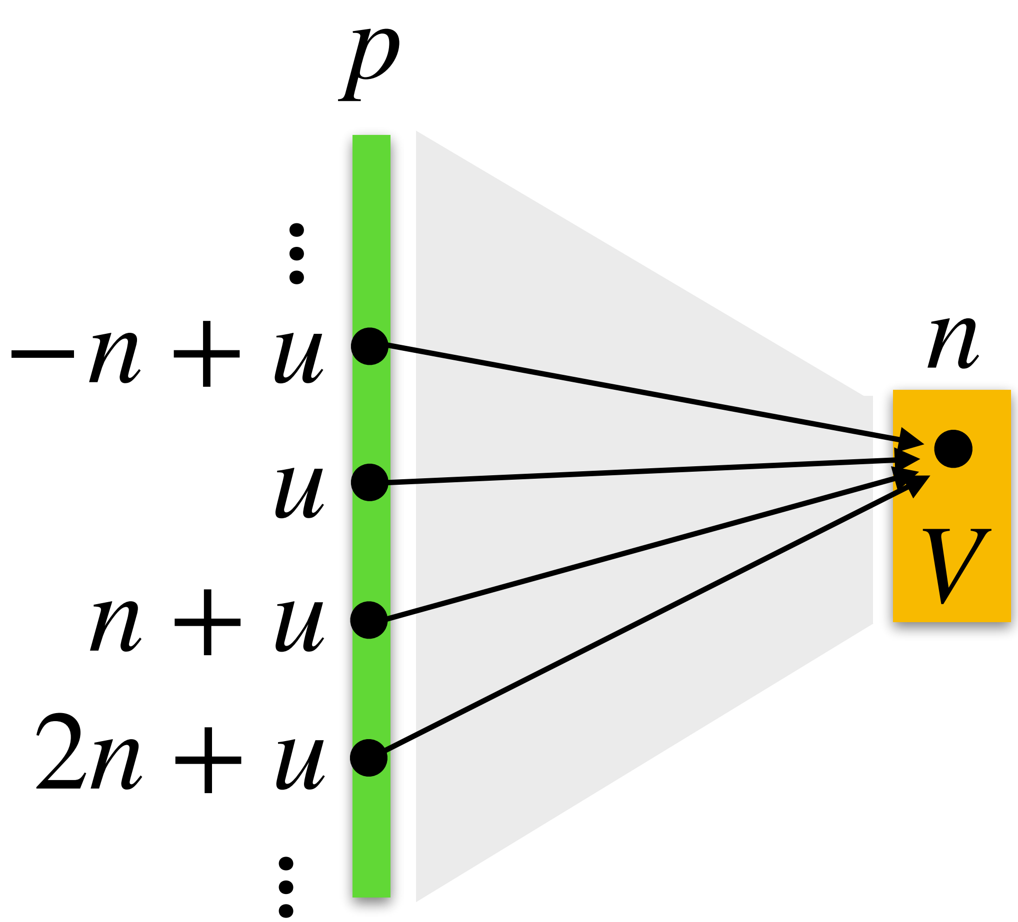
$$((ax + b) \mod p) \mod n$$

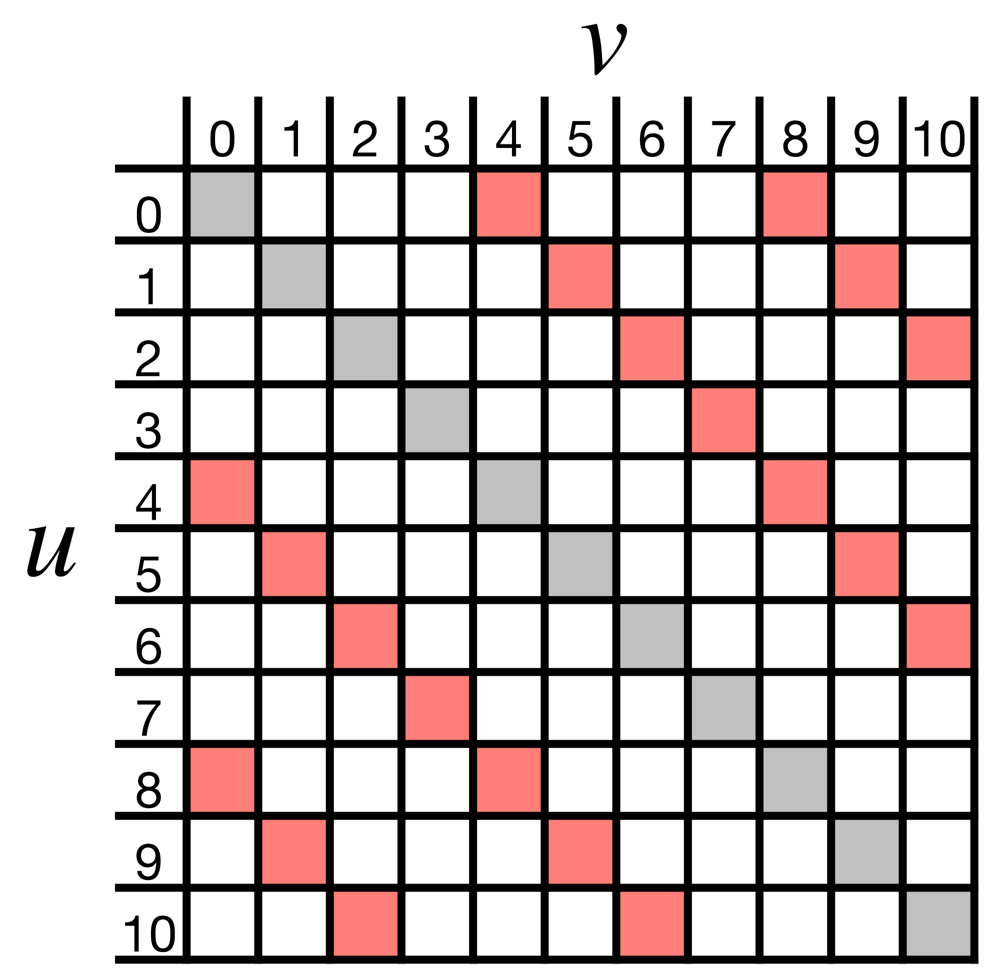


Last concern:

collisions from **final mod** n

Universal hashing

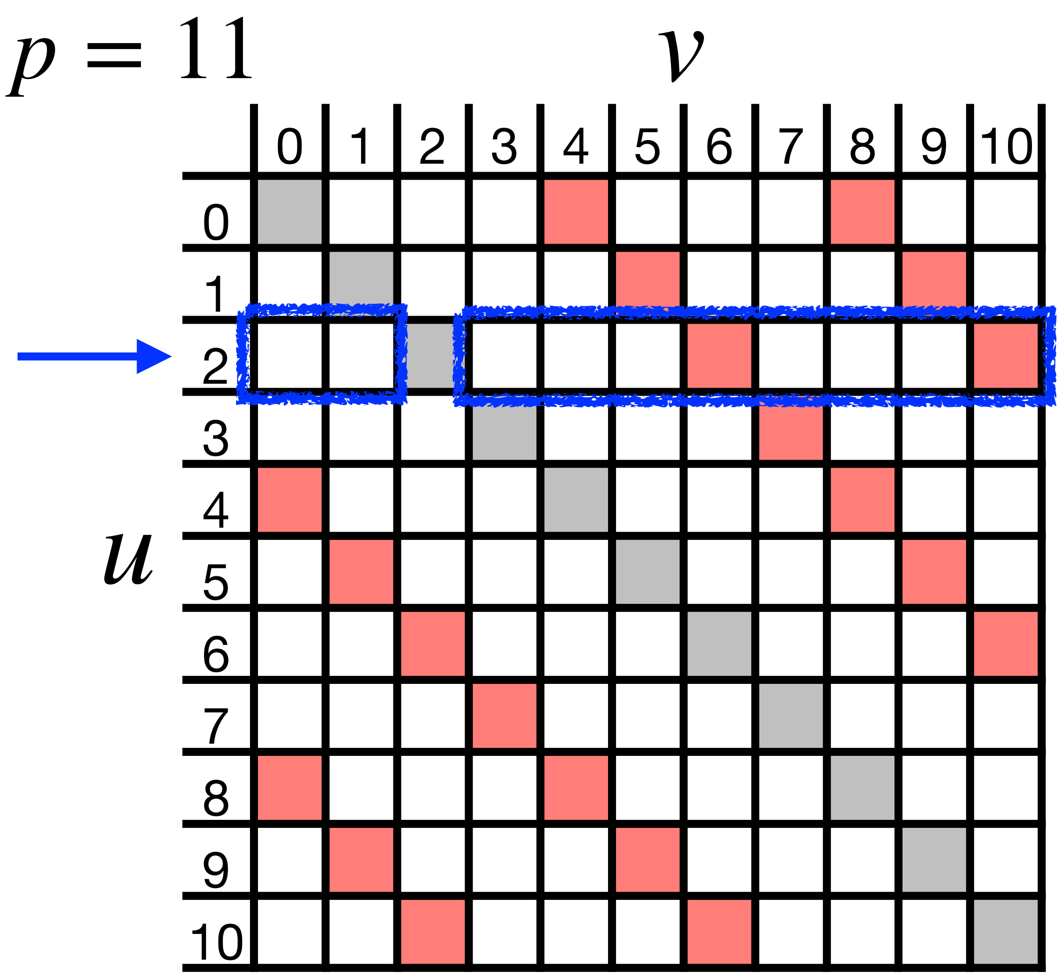




Taking a number u in the prime field, the others $\pm zn$ are its colliders w/r/t V

For $p = 11$ & $n = 4$, 20 out of 110 u, v pairs collide (red squares)

Universal hashing



For given u, number of possible v's ($u \neq v$) is $p - 1$, all equally likely

At most $\lceil p/n \rceil - 1$ choices are collisions

$$\Pr\left(h_{a,b}(x_1) = h_{a,b}(x_2)\right) \leq \frac{\lceil p/n \rceil - 1}{p - 1} \leq \frac{(p-1)/n}{p-1} = \frac{1}{n}$$

2-universality ✅