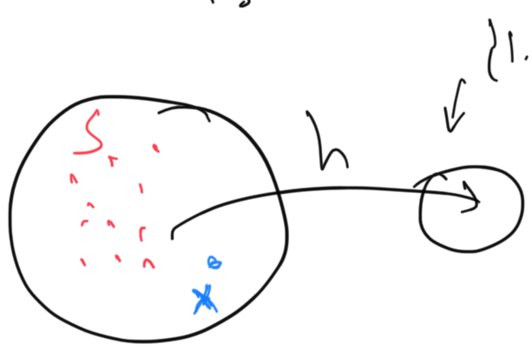
Lecture #3 of AA   1/19/21

- HW 1, due next Thu. Clarity 20%.
- OH.
- Scribe sign-up. 2/lecture.

Dictionary, & Hashing

Dict: prepro. $S \subset U$ s.t. answer quick "is $x \in S$?".



$h: U \to [m]$.

Def: collision $x \neq y$ with $h(x) = h(y)$.

Ideal: enough h s.t. $\forall y \in S$, $\forall x \in U$ $x \neq y \Rightarrow h(x) \neq h(y)$.

Sol Ø: a table of size m, storing whether $i \in \{h(y) \mid y \in S\}$.

Query x check $h(x)$

Issue: very hard to construct eff. such h.f. $\underline{h}$.

e.g.: h: $S = \{i_1 \ldots i_s\}$

$h(i_1) = 1$
$h(i_2) = 2$
$\vdots$

$h(\forall \notin S) = s+1$.

hard to compute $h(x) = ?$

Construct h.f. h trade-offs

- computation/evaluation ti[me] $h(x)$
- how "good" it is for D[S] in terms of collis[ions]

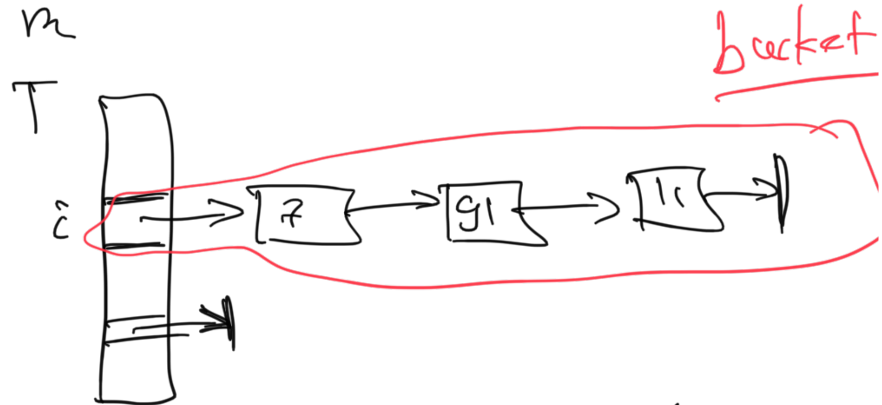
Choose h randomly, allow some collisions.

for query x, $y \in S$, $y \neq x$ $h(x) = h(y)$.

Sol 1: table T with a list of indexes that map into every index.

$T[i]$ stores linked-list of all $y \in S$ s.t. $h(y) = i$.

$i = 1 .. m$



Query alg: @ query x, look up the bucket $h(x)$, and check if x is in the list

Query time? $=$ size of a bucket $+$ time to compute $h(x)$

What hash function?

Remark: [Knuth]: $h(x) = \left\lfloor \left\{ \frac{\sqrt{5}-1}{2} \cdot x \right\} m \right\rfloor$ golden

$U = \{1, \dots |U|\}$. frac

NOT A GOOD HASH FUNC.

$\{7.3\} = .$

Use completely random h.f.

$\mathcal{H} = \{$ all functions $h : U \to [m] \}$.

$|\mathcal{H}| = m^{|U|}$.

$h \in \mathcal{H}$ at random.

Size of bucket?

$= C_x \triangleq$ # elements in S s.t. $h(x) = h(y)$.

C_x ← collision count for x.

$$\mathbb{E}_h[\text{size of bucket}] = \mathbb{E}_h[C_x]$$

except x if $x \in S$

$$= \mathbb{E}_h\Big[\sum_{i \in S} \underbrace{\mathbb{1}[h(i) = h(x)]}_{\text{indicator var} = \begin{cases} 1, & h \\ 0, & o \end{cases}}\Big]$$

$$= \sum_{i \in S} \underbrace{\mathbb{E}_h\big[\mathbb{1}[h(i) = h(x)]\big]}_{1/m \text{ if } i \neq x.}$$

$$= |S| \cdot 1/m \qquad \boxed{n = |S|}$$

$$= n/m.$$

OK to set $m = \Theta(n)$, e.g. $m = n$.

E_h[size of bucket] = 1.

Q. r. t. = $O(1)$ + time to comp. h.
(in expected)

__Question:__ how large is the biggest bucket?

$$\Theta\left(\frac{\lg n}{\lg \lg n}\right).$$

__Question:__ random h.f. h? choose / store?

Actually OK to use h.f. $h \in \mathcal{H}$ with "less" randomness

__Def:__ family $\mathcal{H}$ is __universal__ iff

$$\forall\, x \neq y,\ x, y \in U:\ \Pr_{h \in \mathcal{H}}[h(x) = h(y)] \leq$$

Universal family $\mathcal{H}$ is enough in proof from abo

__Fact:__ $\exists$ univ. hash function fam. $\mathcal{H}$,

$$\lg |\mathcal{H}| = O(\lg U). \qquad |\mathcal{H}| = U^{O(}$$

$\Rightarrow$ describe h using $O(\lg U)$ bits.

<u>Def:</u> $\mathcal{H}$ is <u>α-almost-universal</u> if

$$\forall x \neq y : \Pr_h [h(x) = h(y)] \leq \alpha/m$$

<u>Claim:</u> in the proof above, get

$$\mathbb{E}_h [\text{bucket size, modulo } x] \leq \alpha \cdot n/m.$$

<u>Example:</u> [Dietzfelbinger et al '97]:

$a \in [U]$ at random, odd.

$$h_a(x) = \lfloor ((a \cdot x) \% |U|) \cdot \frac{m}{|U|} \rfloor.$$

$$\mathcal{H} = \{ h_a, \ a \in [U] \text{ odd} \}.$$

<u>Fact:</u> is 2-almost-univ.]

<u>Lemma:</u> can solve Dict. problem, using:

- $O(n)$ space
- $O(1)$ expected q. t.

<u>Pf:</u> $m = n$, table take space:

$$O(m+n) = O(n).$$

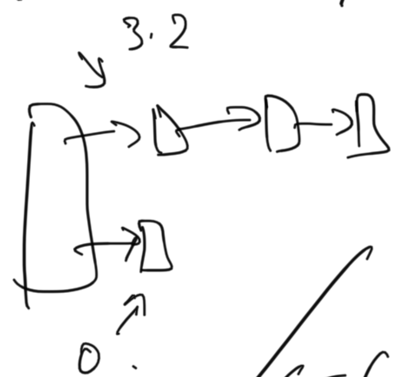
h.f. description is

$\mathbb{E}$[size of bucket] $\le 1 + n/m$.

$O(\lg u)$.

Perfect Hashing:

Goal: $O(1)$ run-time deterministically.



$$C \triangleq \sum_{x \in S} C_x.$$

C_x: # collisions with $y \in S$, $y \neq x$.

$C = 6.$

Ideally: $C = 0.$

$\Rightarrow$ size of bucket $\le 1.$

$$\mathbb{E}[C] = \mathbb{E}_h\Big[\sum_{x \in S} C_x\Big] = \sum_{x \in S} \mathbb{E}_h[C_x]$$

$$= \sum_{x \in S} n/m = \frac{n^2}{m}$$

Suppose $m = 4n^2$. $\Rightarrow$

$\mathbb{E}[C] = 1/4.$

By Markov bound:

$$\Pr[C \ge 4\mathbb{E}[C]] \le \frac{\mathbb{E}[C]}{4\mathbb{E}[C]} = 1/4$$

$\Rightarrow$ with prob $\geq 3/4$, we have:

$$C < 4\cdot\mathbb{E}[C] = 1.$$

$$\Rightarrow C = 0.$$

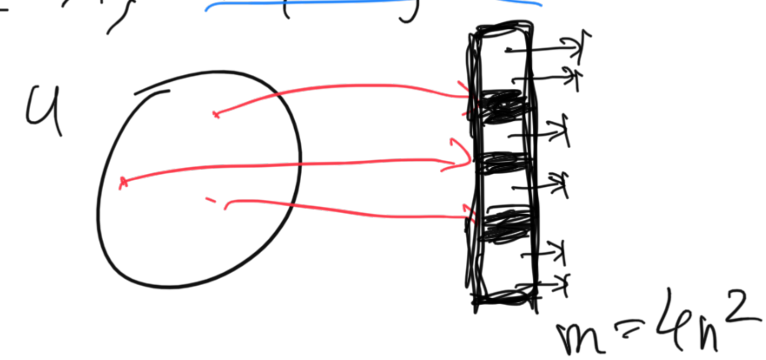
<u>Corollary:</u> can solve Dict. prob.:

- $O(m) = O(n^2)$ space
- $O(1)$ query time (det.).

pf:

- $m = 4n^2$.
- build a hash table using a random $h \in \mathcal{H}$
- compute C. (just check if $\exists$ a collision)
- if $C \geq 1$, try again.

d-a-a.



$\mathbb{E}$[# tries in preproc. algo] =

$$= 1\cdot 1 + 1/4 + (1/4)^2 + (1/4)^3 + \ldots$$

$$= \tfrac{3}{4}\cdot 1 + \tfrac{1}{4}\cdot \tfrac{3}{4}\cdot 2 + \left(\tfrac{1}{4}\right)^2\cdot \tfrac{3}{4}\cdot 3 + \ldots$$

$$= \frac{1}{1-1/4} = 4/3.$$

1-14

Construction time = $O(n^2)$ in expectation

? Choose $h \in \mathcal{H}$ until no collisions

1) $\mathcal{H}$ = all.

$Pr[C=0] > 0$

$\to$ exponential small in n if $m = O(n)$.

tries $\approx 1/Pr[C] \approx \exp(\Theta(n))$.

$\Rightarrow$ $\Theta(n)$ of bits to describe hash func.

2) $\mathcal{H}$ = universal.

$\cancel{Pr[C=0] \overset{??}{>} 0}$ if $m = O(n)$