Lecture 8: Hashing

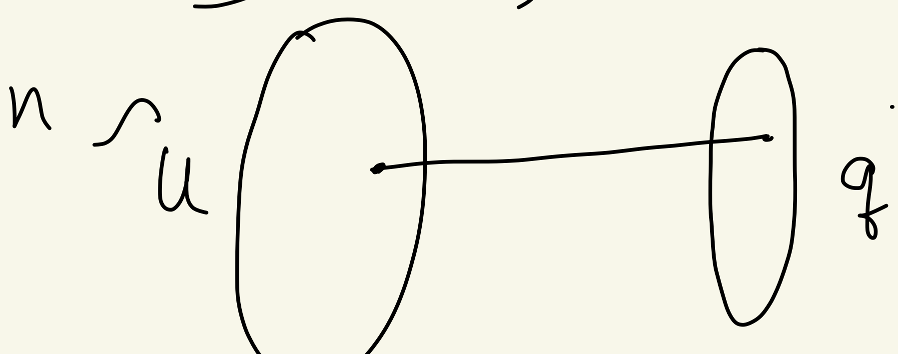
Web Caching

Problem: remember recently accessed web pages

- may need to spread it over multiple machines
- need to quickly determine which server has the page (if any).

Universe $\mathcal{U}$. $|\mathcal{U}| = n$

$S \subseteq \mathcal{U}$, $|S| = m \ll n$



<u>Insert</u>(x)

Delete(x)

Query(x)

Need: "oblivious" to the specific set S

$$h : U \to [q] = \{0, 1, 2, \dots, q-1\}$$

For any h, there's going to be an $i \in [q]$ s.t. $|h^{-1}(i)| \geq \frac{n}{q}$

$$h^{-1}(i) := \{x \in U \mid h(x) = i\}$$

Choose a random hash function.

For every x indep, $h(x) = i$ w.p. $1/q$ $\forall i \in \{1, \ldots, q\}$

Let's reason about the # elements in S hashed into a single cache location i.

X_i = # elements hashed into location i.

$\mathbb{E} X_i$?

$$X_{ij} = \begin{cases} 1 & \text{if } j^{th} \text{ element in } S \text{ is mapped into loc } i \\ 0 & \text{o/w} \end{cases}$$

$$X_i = \sum_{j \in S} X_{i,j}$$

$$\mathbb{E} X_{ij} = \Pr[X_{ij} = 1] = \frac{1}{q}.$$

$$\mathbb{E}[X_i] = \sum_{j \in S} \mathbb{E}[X_{ij}] = \frac{|S|}{q} = \frac{m}{q}$$

If $m \leq q$, then $\mathbb{E} X_i \leq 1$

"high probability estimate" of # elements hashed to loc i

$$\mathrm{Var}(X_i) = \mathrm{Var}\Big(\sum_{j \in S} X_{ij}\Big)$$

$$= \sum_{j \in S} \mathrm{Var}(X_{ij})$$

$$= \sum_{j \in S} \frac{1}{q}\Big(1 - \frac{1}{q}\Big)$$

$$\leq \sum_{j \in S} 1/q = \frac{m}{q} \leq 1$$

Chebyshev:

$$\sigma = \sqrt{\mathbb{E}(X_i - \mathbb{E}X_i)^2} = \sqrt{\mathrm{Var}(X_i)}$$

$$\Pr\left[\,|X_i - \underline{\underline{1}}| \geq t\sigma\,\right] \leq 1/t^2$$

$$\underbrace{\Pr\left[\,|X_i - 1| \geq t\,\right]}_{B_i} \leq 1/t^2$$

Union Bound

$$\Pr\left[B_0 \vee B_1 \vee \dots \vee B_{q-1}\right]$$

$$\leq \Pr[B_0] + \Pr[B_1] + \dots + \Pr[B_{q-1}]$$

$$\leq \frac{q}{t^2} \leq \frac{1}{2} \qquad \left(\text{So } t \sim \sqrt{2q} \text{ suffices}\right)$$

Chernoff:

$$X_i = \sum_{j \in S} X_{ij}$$

$$\boxed{\Pr[X \geq (1+\delta)\cdot\mu] \leq e^{-\frac{\delta^2}{2+\delta}\cdot\mu}}$$

[Since $2+\delta \leq 2\delta$

$\frac{1}{2+\delta} \geq \frac{1}{2\delta}$

$\Rightarrow -\frac{1}{2+\delta} \leq -\frac{1}{2\delta}$]

$$\leq e^{-\frac{\delta^2}{2\delta}\mu} = e^{-\frac{\delta\mu}{2}} \text{ for } \delta > 1$$

for large δ, exponent is linear in δ

$$\Pr[X_i \geq \underline{1}(1 + 2(\log q + 100))$$

$$\leq e^{-\frac{\delta \cdot \mu}{2}} = e^{-\frac{2(\log q + 100)}{2}}$$

$$= e^{-\log(q)} \cdot e^{-100}$$

$$\leq \frac{1}{q} \cdot e^{-100}$$

B_i: event that

$$X_i \geq (1 + 2(\log q + 100))$$

$$\Pr[B_0 \vee \ldots \vee B_{q-1}]$$

$$\leq q \cdot \frac{1}{q} \cdot e^{-100} = e^{-100}$$

w.p. $1-e^{-100}$,
all locations must
have at most
$1+2\log q+200$
elements.

Idea: Choose h randomly from a small hash family so that

1) no lookup table needed
2) can efficiently compute h at any element
3) spreads out load evenly across hash locs

$\mathcal{H} \subseteq \{h : U \to [q]\}$

Def: $\mathcal{H}$ is a "2-universal" hash family if for every $x, y \in U$, $x \neq y$:

$$\Pr_{h \sim \mathcal{H}}[h(x) = h(y)] \leq \frac{1}{q}$$

Lemma: Fix $x \in S$.

Let X be the random variable that counts the # elements $y \in S$, s.t. $h(y) = h(x)$.

Then $\mathbb{E}X \leq 1 + \frac{m-1}{q} \leq 2$

(if $m \leq q$)

Pf:

$$I_y = \begin{cases} 1 & h(y) = h(x) \\ 0 & o/w \end{cases}$$

$$\boxed{X = \sum_{\substack{y \in S \\ y \neq x}} I_y}$$

$$\mathbb{E}X = \sum_{y \in S, y \neq x} \mathbb{E} I_y = \sum_{y \in S, y \neq x} \mathbb{P}r[h(y) = h(x)]$$

$$\leq \sum_{\substack{y \in S, \\ y \neq x}} 1/q = \frac{m-1}{q}$$

Construction of a 2-universal hash family

1. Choose a prime $p \in \{|U|, \ldots, 2|U|\}$

Bertrand's Postulate ?

2. $f_{a,b}(\underline{x}) = \underline{a}x + \underline{b} \bmod p$

3. $h_{a,b}(x) = f_{a,b}(x) \bmod q$

$$\mathcal{H} = \left\{ h_{a,b} \;\middle|\; \begin{array}{l} a \in \{1, \ldots, p-1\} \\ b \in \{0, \ldots, p-1\} \end{array} \right\}$$