Note Title 1/13/2016

Statistical Principals

Hashing

- Birthday Paradox
- Coupon Collectors Problem
- Central Limit Theorem

Data $X = \{x_1, x_2, \ldots, x_n\}$

$x_i \sim D(\theta)$

iid — θ: parameters

- independently and identically distributed

Big Data

complex / accurate estimates

combining many small observations

Hash functions

$h_a \in \{ \quad \}$

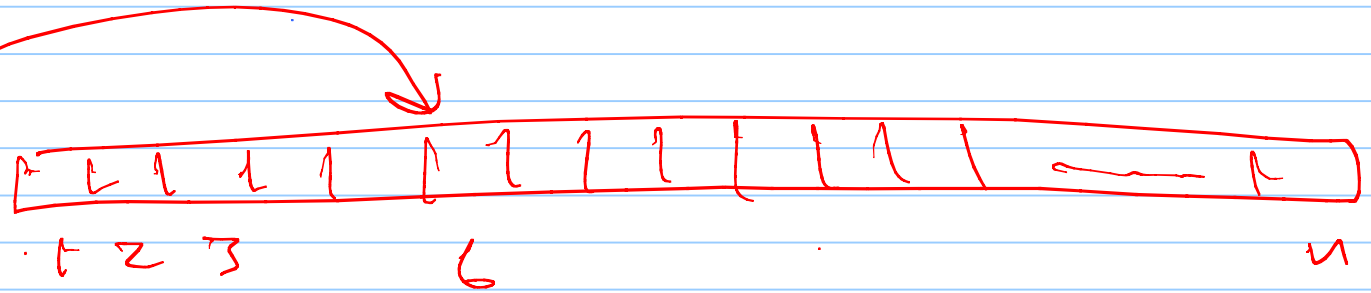
$h : [m] \to [n]$

all IP addresses

$\{1, 2, 3, \ldots, n\}$

$x \in [m]$

$h(x)$



$\Pr[h_a(x) = h_a(x')] = \frac{1}{n}$ $\quad x, x' \in [m]$

SHA-1 (Σ^*)

↳ $n = 2^{160}$

Σ = string of bits

$\{0,1\}^*$

add salt : string a

input x

SHA-1 (concat (a, x))

Multiplicative Hashing

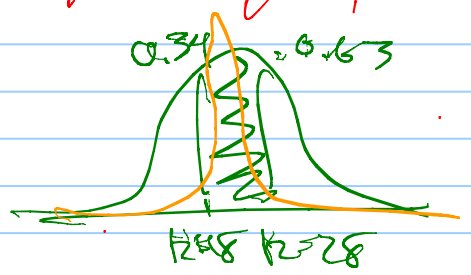
$h_a(x) = \lfloor n \cdot frac(x \cdot a) \rfloor$

Modular Hashing : $h(x) = x \bmod n$

Birthday Paradox

$h : [\text{person}] \to [365]$

Jan :
Feb :
Mar :
Apr :
May :
June:
Jul :
Aug :
Sep :
Oct :
Nov :
Dec :



$\Pr[\text{collision } k \in (18, 28)]$

1st 0.3 person

2nd person

$\frac{1}{365}$ First collision

$k \approx \sqrt{2n}$

k people

$\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right) \cdots \left(\frac{n-k}{n}\right)$ $\binom{k}{2}$ pairs $\approx \frac{k^2}{2} = \frac{k(k-1)}{2}$

$\Pr[\text{no col}] = \left(1 - \frac{1}{365}\right)^{\frac{k^2}{2}}$

$k = 23$

$(0.997)^{253} = 0.467$

Jan :
Feb :
Mar :
Apr :
May :
June:
Jul :
Aug :
Sep :
Oct :
Nov :
Dec :

$n^2 = 1424$

r_0 = # tracks before seeing dth distinct coupon

$n \ln(n) = 36$

$t_d = r_d - r_{d-1} = r_d - r_{i-1}$

tracks between $(d-1)$th dist coupon and dth

$r_n = \sum_{d=1}^{n} t_d$

$\Pr[\text{selected}] = \frac{n-d+1}{n}$ $\mathbb{E}[t_d] = \frac{n}{n-d+1}$

$\mathbb{E}[r_n] = \sum_{d=1}^{n} \frac{n}{n-d+1} = n\left(\sum_{d=1}^{n} \frac{1}{d}\right) = nH_n = n(\ln n + \gamma)$

Coupon Collectors

$= \gamma + \ln(n) + o(1)$

≈ 0.577

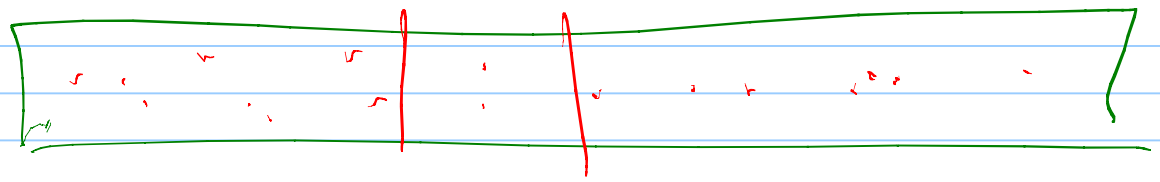
$h^x \longrightarrow [n]$ How many draws to get all at least n draws

---

n events $\rightsquigarrow \{p_1, p_2, \dots, p_n\}$ $\sum_i p_i = 1$

$\mathbb{E}[X] = \frac{1}{p_*}(\gamma + \ln(n))$ $\quad p_* = \min_i p_i$

$\varepsilon = (0,1)$ $\Rightarrow \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$



Central Limit Thm / Chernoff-Hoeffding Bound

Random Variables $X_1 \; X_2 \; \ldots \; X_n$

$-\Delta \le X_i \le \Delta$

$-125 \qquad 125$

$A = \frac{1}{n} \sum_{i=1}^{n=100} X_i$

$$Pr\left[|A - E[A]| > \alpha\right] \le 2 \exp\left(\frac{-n\alpha^2}{2\Delta^2}\right)$$

10

$\exp(x) = e^x$