HW4 due Tue Mar 1

Randomized algo's with prob. of error

Las Vegas algo: Always correct, probably fast

Monte Carlo algo: Always fast, probably good/correct

← Area = 1

$\Pr[p \in B] = \text{Area}(B)$

Est area = Fraction of samples in B

Monte Carlo: errorrate δ confidence $1-\delta$

$$\Pr[\text{wrong/bad}] < \delta$$

Set membership: Insert Query with error δ

one-sided error

$x \in S \Rightarrow$ YES

$x \notin S \Rightarrow$ NO with prob $1-\delta$

YES w/prob $< \delta$

$\boxed{\text{FILTER}}$

ZATOCODING '57 BloomFilter '70

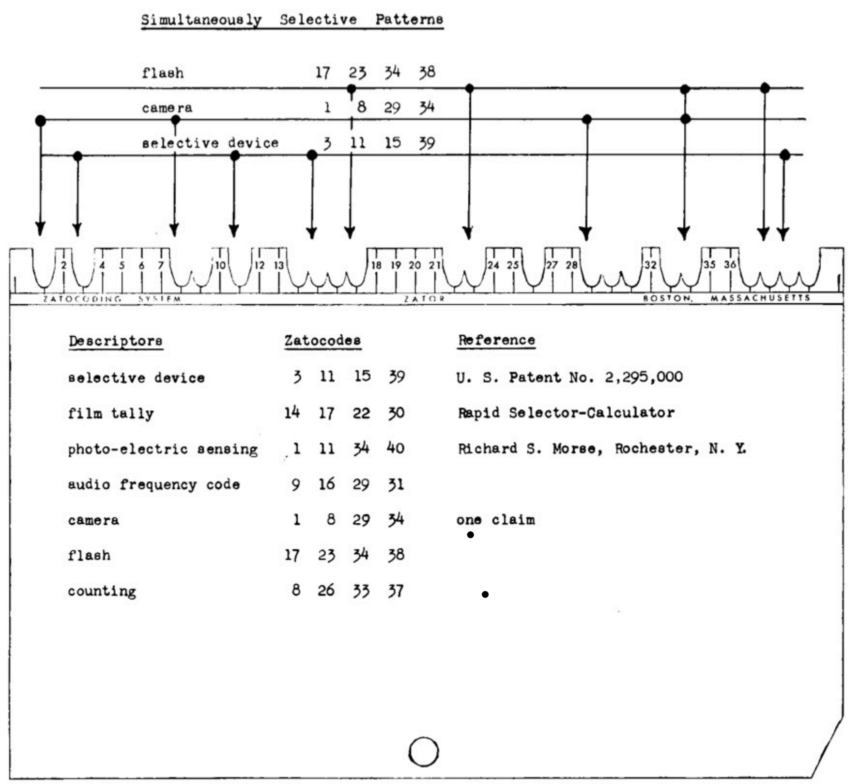


Fig. 1. ZATOCODING, illustrated with a 5 by 8 inch edge notched Zatocard for the tally. Note that the random Zatocode patterns in the field overlap and intermingle. Selection on the combination of three descriptors, "flash," "camera," and "selective device," is according to the inclusion of the pattern of arrows into the pattern of notches in the coding field. Zatocards are sorted by the selector shown in Figure 2.

Bloom filter: bits $B[0..m-1]$

$+\ k$ hash functions $h_i : \mathcal{U} \to [m]$

ideal random

Insert(x):

for $i \leftarrow 1$ to k
$\quad B[h_i(x)] \leftarrow 1$

Member?(x):

for $i \leftarrow 1$ to k
$\quad$ if $B[h_i(x)] = 0$
$\quad\quad$ FALSE
TRUE

$$\Pr[h_i(x) = j] = 1/m$$

$$\Pr[\text{Insert}(x) \text{ does } \underline{\text{not}} \text{ set } B[j] \leftarrow 1] = \left(1 - \frac{1}{m}\right)^k$$

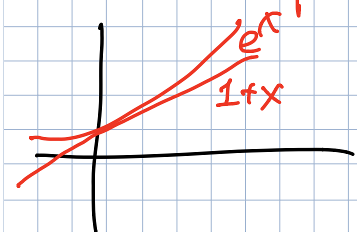
After n Inserts:

$$\underbrace{\Pr[B[j] = 0]}_{p} = \left(1 - \frac{1}{m}\right)^{kn} < \boxed{e^{-kn/m}}$$

WMU ≠

$e^x \geq 1 + x$

$$\Pr[\text{False positive}] = (1 - p)^k \quad \text{(sort of)}$$



$$\boxed{\delta = \left(1 - e^{-kn/m}\right)^k}$$

$$\ln \delta = \ln(1-p)^k = k \ln(1-p) = -\frac{m}{n} \ln p \ln(1-p)$$

max at $\boxed{p = 1/2} \Rightarrow \boxed{k = \frac{m}{n} \cdot \ln 2}$

$$\delta = \left(\tfrac{1}{2}\right)^{\frac{m}{n}\ln 2} \approx (0.6185)^{m/n}$$

$$\boxed{m = \frac{\lg(1/\delta)}{\ln 2} \cdot n} \Rightarrow \text{error prob } \delta$$

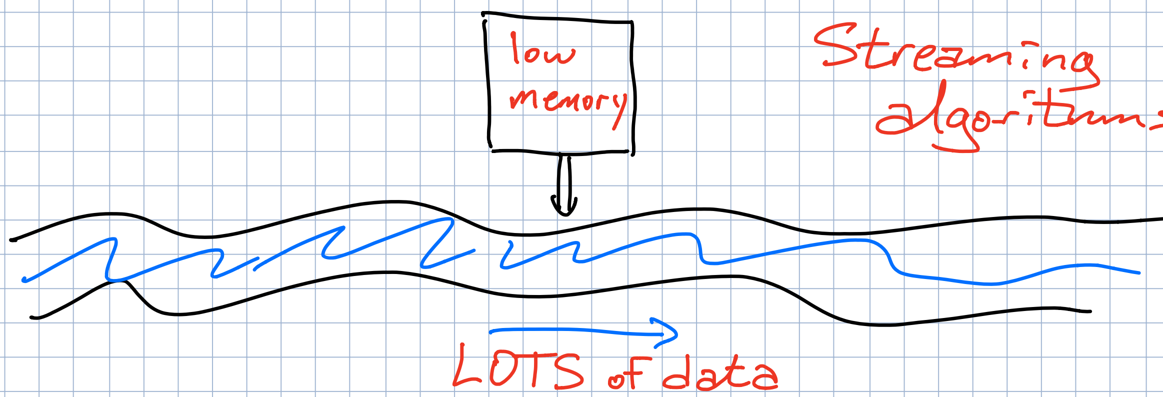
$$= O(n \cdot \log(1/\delta))$$

$\delta = 1\% \Rightarrow 10n$ bits $\quad k = 7$

$m = 32n \Rightarrow k = 22 \quad \delta \approx 2 \cdot 10^{-7}$

Mitzenmacher
Broder

Streaming algorithms



"How many times have I seen item x?"

Count Min Sketch

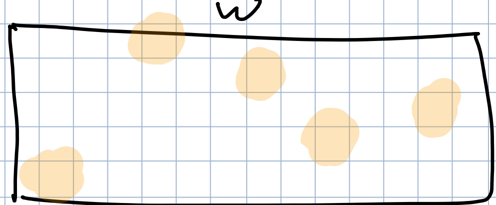
$w \times d$ array of integers
width depth

d universal hash functions

$h_i : \mathcal{U} \to [w]$

w

d



Process(x):

For $i \leftarrow 1$ to d
  $\mathrm{Count}[i, h_i(x)]$++

Estimate(x):

$\min_i \mathrm{Count}[i, h_i(x)]$

If we choose w and d correctly

total length of stream so far

$$\Pr[\mathrm{Estimate}(x) > \mathrm{freq}(x) + \varepsilon \cdot N] \leq \delta$$

$w = \left\lceil \frac{e}{\varepsilon} \right\rceil \qquad d = \lceil \ln(1/\delta) \rceil$

$$E[X_{i,x}] = \#\text{collisions with } x \text{ in row } i$$
$$= \sum_{y \neq x} \Pr[h_i(x) = h_i(y)] \cdot Freq(y)$$
$$\leq \frac{N}{w}$$

Markov's inequality.

$$\Pr[X_{i,x} > \varepsilon N] \leq \frac{1}{w\varepsilon}$$

$$\Pr[Est > freq(x) + \varepsilon N] \leq \left(\frac{1}{w\varepsilon}\right)^d < \delta$$