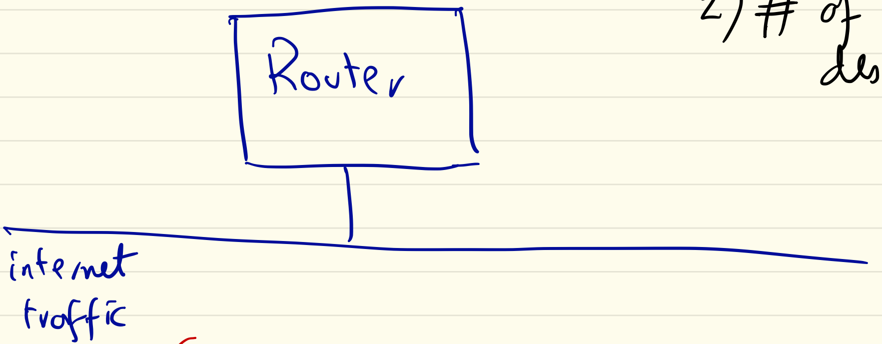
# Streaming Algorithms

Input is a stream

SITUATION:




1) # of packets
2) # of IP-address destination.

each packet (IP-address destination — message)

128 bit address

sees $2^{40}$ packets in a day

Streaming Alg:

INPUT: A stream $s_1, s_2, \ldots, s_n$

of each $s_i \in \{1, \ldots, N\}$

GOAL: Compute function$(s_1, \ldots, s_n)$

statistic

RESTRICTION: 1) Memory available to Alg $\ll n, N$

$\simeq \text{poly}(\log n, \log N)$

2) See the input only once in the order $s_1, \ldots, s_n$

~~Part~~ Streaming:

- Read a book with a little sheet to take notes
- ~~And~~ Compute Word Statistics

- Observe traffic

INPUT: A stream $s_1, \ldots$

[ read a book with words ]

SOL:

ALG: Sampling - "Keep a random sample"

300 m

300 voters who vote 0 or 1

estimate # of voters who ~~are~~ 1

- sample $k$ voters $X_i \in \{0,1\}$
- Output $\frac{1}{k} \sum X_i$

With $k$-samples, the estimate $\frac{1}{k} \sum X_i$

is correct within accuracy $\varepsilon$

with probability $1-\delta$

1) $k \cdot n$ 2) $k$ 3) $k \cdot \log n$ 4) $k \cdot \sqrt{n}$

## Chernoff/ Hoeffding Bound

Suppose $s_1, \ldots s_n$ is a stream of $\{0,1\}$

$p$ fraction of 1s

let $X_1, \ldots X_k \leftarrow$ be uniformly random samples from $s_1, \ldots s_n$

$X_\ell \leftarrow s_{i_\ell}$ (uniformly random)

Output: $\frac{1}{k} \sum X_i$

$$\mathbb{E}\, X_i = \{\text{fraction of 1s in } s_i\} = p$$

estimate, answer

$$\Pr\left[\left|\frac{1}{k}\sum X_i - p\right| > \varepsilon\right] \leq 2e^{-2\varepsilon^2 k}$$

# RESERVOIR SAMPLING

Input : Stream $s_1 \ldots s_n$ ← don't know when stream ends.

GOAL : A uniformly random element from stream.

"Current choice"

reservoir $= s_1$ reservoir$[1..t] \leftarrow s[1..t]$

for $i=2$ to . . . . .

choose random number $r \in \underbrace{\{1 \ldots i\}}$

If $(r=1)$ reservoir $\leftarrow s_i$ If $(r \leq t)$ reservoir$[r] = s_i$

else Ignore $s_i$

Output reservoir

INDUCTIVE HYPOTHESIS

. At end of iteration $i$,

$$\Pr[\text{reservoir} = s_j] = 1/i$$

for $j = 1 \ldots i$

$i = 1$ $\quad \Pr[\text{reservoir} = s_1] = 1$

At iteration $i+1$, $\quad j \in \{1 \ldots i\}$

$$\Pr[\text{reservoir} = s_j] = \left(\begin{array}{c}\text{reservoir} = s_j \\ \text{at iteration } i\end{array}\right) \text{ AND } \left(\begin{array}{c}s_j \text{ NOT} \\ \text{kicked out}\end{array}\right)$$

$$= \frac{1}{i} \cdot \left(1 - \frac{1}{i+1}\right) = \left(\frac{1}{i+1}\right)$$