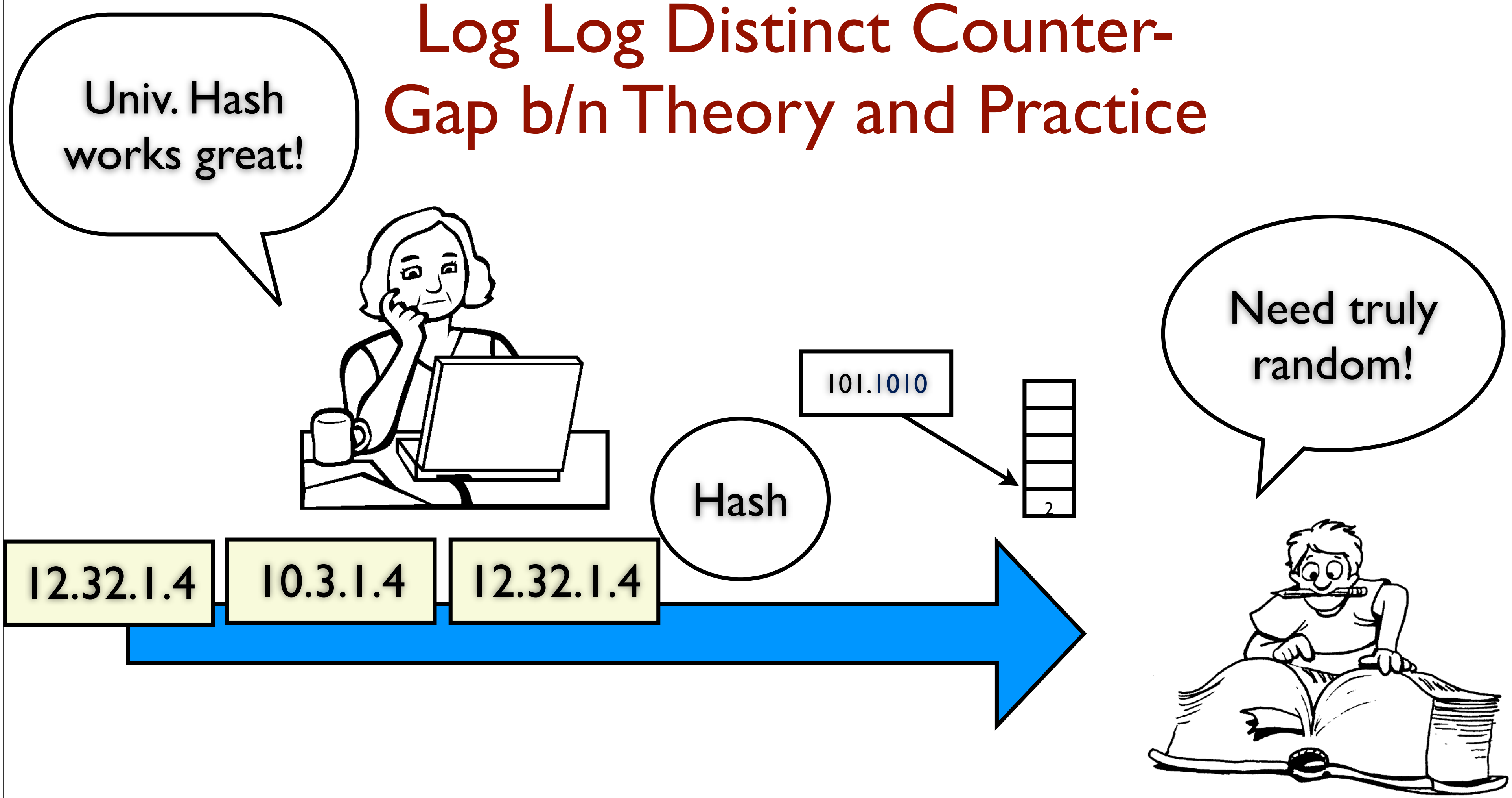
Log Log Distinct Counter- Gap b/n Theory and Practice



*"people have studied constructions of hash functions using limited randomness that work with worst case data, but such constructions are cumbersome and in practice, people use much simpler hash functions."

Random Data + Universal H(x) $\Longleftrightarrow$ Worst Case Data + Truly Random H'(x)

Pf. Block Source Data has 'cond. collision. prob. per block'. Univ. H 'extracts' this to give close to uniform hash values (MV08)

$$\boxed{x_1}\boxed{x_2}\boxed{X_3}$$

$$\left.\begin{array}{l} \mathrm{cp}(X_3 \mid x_1, x_2) < 1/K \\ \mathrm{P}\{H(x_x) = H(x_2)\} < 1/M \end{array}\right\}$$

$Y = H(X_1), H(X_2), H(X_3)$ is ϵ close to Z

and Z has cp $< 3/M^3$ **if** $\boxed{K = O(2MT^2/\epsilon)}$

$*\ cp(X) = \sum_x \mathrm{P}\{X = x\}^2$
collision probability

How much is $\max_{(x_1, x_2, \ldots, x_{i-1}) \in [N]^{i-1}} \mathrm{cp}(X_i \mid x_1, x_2, \ldots, x_{i-1})$

From Real Traffic Traces!

But we'd never have enough examples to measure this for all possible values of

$(x_1, x_2, \ldots, x_{i-1}) \in [N]^{i-1}$

Find $\mathrm{cp}(X_i \mid x_1, x_2, ..., x_{i-1})$ weighted by prob. of seeing $(x_1, x_2, ..., x_{i-1})$ in the trace.

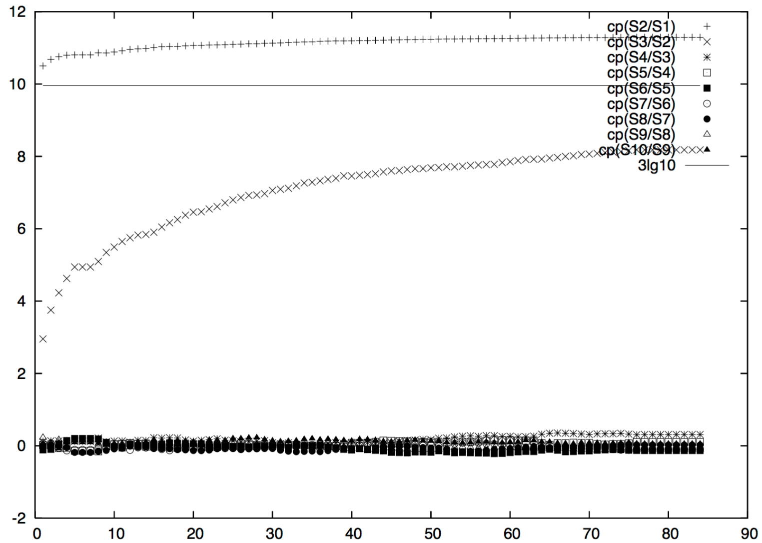
If avg.$_i$ < 1/K² then hash still close to uniform . . .

But this is not scalable, need to keep individual counts for too many $(x_1, x_2, ..., x_{i-1})$

Find $\mathrm{cp}(X_i \mid x_1, x_2, \ldots, x_{i-1})$ weighted by **prob.²** of seeing $(x_1, x_2, \ldots, x_{i-1})$ in the trace.

$$= \frac{\sum_{(x_1,\ldots,x_{i-1})} freq.(x_1,\ldots,x_{i-1})^2}{\sum_{(x_1,\ldots,x_i)} freq.(x_1,\ldots,x_i)^2}$$

easy to sketch accurately and scalably (using `F2 sketches')



Evaluated on a 10 second (280mn pkt) trace, found estimate for S3/S2, need larger trace for S4/S3 onwards . . .

Still <u>need large amounts of data</u> to get good estimates . . . though it can be done efficiently

Challenges . . .

- understanding new concepts (entropy, randomness extractors, etc)

- understanding sophisticated and diverse kinds of worst case analysis for different distinct counters and applications (Linear Probing, Balanced Allocations etc. which were analyzed in the "block-sources paper")

Contributions . . .

- applied “block-source model” based techniques to find how ‘random’ traffic needs to be for LogLog Counter

- explored empirical verification of “Internet traffic is random enough”

- introduced new “average” measures of randomness in traffic, one can be computed efficiently. Both still need a lot of data.

Open questions . . .

- Can the proposed randomness measure be estimated in small space ?
    - "avg. condn. cp weighted by prob. ^{2}"? Yes, using F2 sketches
    - "avg. condn. cp weighted by prob."??

- Can the theoretical proofs of randomness required be adapted to use the new average randomness measures ?
  - "avg. condn. cp weighted by prob."? Yes, though it needs to be less than $1/K^2$ (v/s max. condn. cp $<$ $1/K$), is there a better bound??
  - "avg. condn. cp weighted by prob. ^{2}"??