An optimal cardinality estimation algorithm based on order statistics and its full analysis

Jérémie Lumbroso

Projet ALGORITHMS, INRIA Rocquencourt &
Laboratoire d'Informatique de Paris 6 (UPMC).

AofA'10

July 2, 2010—Vienna

1. PROBLEM STATEMENT

Definition: a stream $\mathcal{S}$,

$$\mathcal{S} = s_1 s_2 \cdots s_N.$$

size of the stream: $|\mathcal{S}| = N$

cardinality (= number of distinct elements): $\|\mathcal{S}\| = n$

For instance,

$$\mathcal{S} = run, sally, run, see, sally, run \qquad |\mathcal{S}| = 6 \qquad \|\mathcal{S}\| = 3.$$

1. PROBLEM STATEMENT

Definition: a stream $\mathcal{S}$,

$$\mathcal{S} = s_1 s_2 \cdots s_N.$$

size of the stream: $|\mathcal{S}| = N$

cardinality (= number of distinct elements): $\|\mathcal{S}\| = n$

For instance,

$$\mathcal{S} = run, sally, run, see, sally, run \qquad |\mathcal{S}| = 6 \qquad \|\mathcal{S}\| = 3.$$

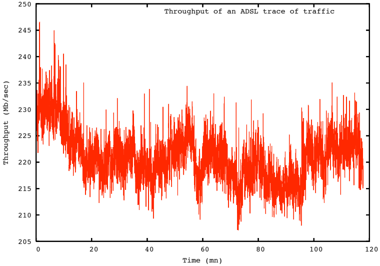
Problem: Estimate cardinality of $\mathcal{S}$ so large that it cannot be stored.

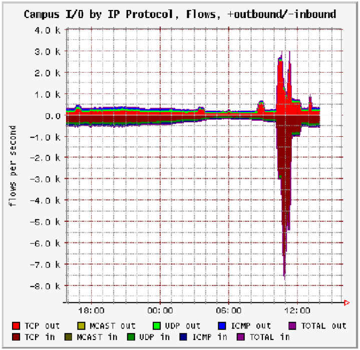
Constraints

- very little processing memory
- on the fly (single pass + simple main loop)
- no statistical hypothesis
- accuracy within a few percentiles

MOTIVATION

- **Network security**:
  detect attacks (denial of service), or the spreading of worms/spam,...





- **Data mining**: document classification, ...
- **Databases**: query optimization
- **Distributed**: censor networks

Bibliographic context

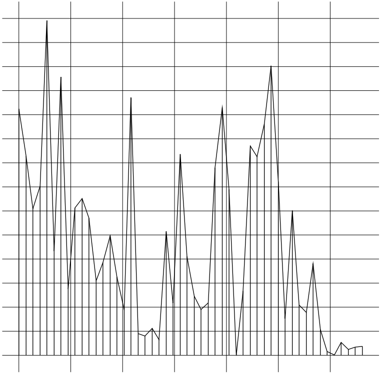
1. Algorithms based on pattern observation
   - Flajolet and Martin, 1985, *Probabilistic Counting*
   - Durand and Fl., 2003, *Loglog* ; Fl. and al., 2007, *Hyperloglog*
2. Algorithms based on order statistics
   - Giroire, 2003-2006, thèse P6
3. Complexity results
   - Alon, Matias, Szegedy, 1996, Frequency moments
   - Chassaing and Gerin, 2006, Theoretical optimality of using the minimum

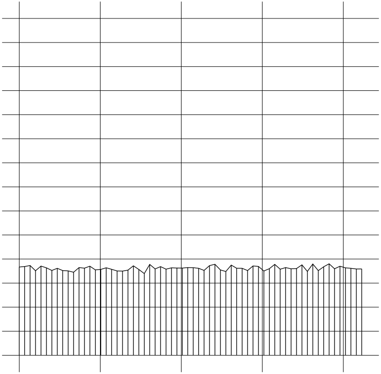
2. THE MODEL

Definition: a hash function h is defined as

$$h : \mathcal{A}^* \longrightarrow [0, 1].$$

Main idea. With "good enough" hash functions, our data is uniformized.





Definition: an observable = function of the underlying hash set (i.e.: a function not sensitive to repetitions)

Example: minimum

- $\min\{1,2,3\} = \min\{1,\dots,1,2,\dots,2,3,\dots,3\} = 1$

With:

- hash functions that uniformize the data
- observables : functions of underlying hash set

process data → study **n i.i.d. random variables in $[0,1]$**.

3. ORDER STATISTIC of rank 1

M := minimum of n random variables

$$\mathbb{P}_n[M \in [x, x + \mathrm{d}x]] = n(1-x)^{n-1}\mathrm{d}x \qquad (1)$$

hence

$$\mathbb{E}_n[M] = \int_0^1 x \cdot n(1-x)^{n-1}\mathrm{d}x = \boxed{\frac{1}{n+1}}. \qquad (2)$$

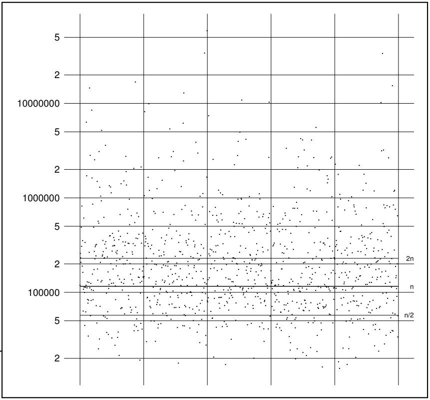
Advantages:

- computable in one pass
- computable with a single register
- $\mathbb{E}_n[M] = \frac{1}{n+1}$.

Disadvantages:

- minimum oscillates a lot, indeed $\sigma_n[M] = \frac{1}{n+1}$
- function $x \mapsto \frac{1}{x}$ diverges at 0.

$$\mathbb{P}_n\left[M \leqslant \frac{t}{n}\right] \sim 1 - \exp(-t)$$



STOCHASTIC AVERAGING

How to cheaply repeat the estimation (to average)?

Idea. Make m copies of the $[0, 1]$ interval, and distribute the hashed values on these m intervals.

An extra condition: a **given element** must always be attributed to the **same interval**.

Core Algorithm

Parameter: m control parameter
Input: a stream $S = (s_1, \ldots, s_N)$

---

initialize m registers M_1 through M_m to 1

forall $x \in S$ **do**

$A := h(x)$ {hash x, with $h(x) \in (0,1)$}
$j := \lfloor mA \rfloor + 1$ {index of the substream assigned to x}
$M_j := \min(M_j, mA - \lfloor mA \rfloor)$ {update minimum of j-th substream}

return $\mathcal{Z}^* = m \cdot \frac{(m-1)}{M_1 + \ldots + M_m}$

4. ANALYSIS of the CORE algorithm

$$\mathcal{Z}^* = m \cdot \frac{(m-1)}{M_1 + \ldots + M_m}$$

Configuration $\mathcal{C}$ defined by:

- allocation of n RVs in m bins (stochastic averaging)
- minimum of each bin

$$\mathbb{P}_n[\mathcal{C}] = \frac{1}{m^n}\binom{n}{n_1, \ldots, n_m} \prod_{j=1}^{m} n_j (1-x_j)^{n_j - 1} \mathrm{d}x_j. \qquad (3)$$

[Interm.] Lemma. The r-th moment of random variable $\mathcal{Z}^\star$ is given by

$$\mathbb{E}_n[(\mathcal{Z}^\star)^r] = \bullet \int_{[0, \frac{n}{m}]^m} \left(1 - \frac{1}{n}\sum_{j=1}^{m} t_j\right)^{n-m} \frac{\mathrm{d}t_1 \cdots \mathrm{d}t_m}{(t_1 + \ldots + t_m)^r} \qquad (4)$$

Proof.

1. sum (3) over all configurations
2. integrate over all possible minima: $x_j \in [0, 1]$
3. rescaling ($x_j = m/n \cdot t_j$) and algebraic manipulations.

Calculating the multi-dimensional parametered integral

$$\mathbb{E}_n[(\mathcal{Z}^\star)^r] = \bullet \int_{[0,\frac{n}{m}]^m} \left(1 - \frac{1}{n}\sum_{j=1}^{m} t_j\right)^{n-m} \frac{\mathrm{d}t_1 \cdots \mathrm{d}t_m}{(t_1 + \ldots + t_m)^r}$$

Laplace method:

1. split integral into

$$I_C = \left[0, \frac{\delta(n)}{m}\right]^m \qquad \text{and} \qquad I_T = \left[0, \frac{n}{m}\right]^m \setminus I_C$$

2. on I_C, use $(1 - \frac{1}{n}\sum) \sim \exp(-\sum)$
3. show I_T is negligible

+ use integral representation of Gamma function on integers

$$\int_0^\infty \mathrm{e}^{-ay} a^{r-1} \mathrm{d}a = \frac{(r-1)!}{y^r}.$$

A) UNBIASED and ACCURATE

Theorem 1: $\mathcal{Z}^\star$ is asymptotically unbiased, in the sense that

$$\mathbb{E}_n[\mathcal{Z}^\star] = n(1 + o(1)). \tag{5}$$

Theorem 2: The precision of estimator $\mathcal{Z}^\star$, expressed in terms of standard error, satisfies

$$\frac{\sigma_n[\mathcal{Z}^\star]}{n} \sim \frac{1}{\sqrt{m-2}}. \tag{6}$$

B) LIMIT DISTRIBUTION

Let $S := M_1 + \ldots + M_m$, where the M_j are interdependent, the Laplace transform,

$$\mathbb{E}\left[\mathrm{e}^{-wS}\right] \sim \left(\int_0^\infty \mathrm{e}^{-t}\mathrm{e}^{-w\frac{m}{n}t}\mathrm{d}t\right)^m = \left(\mathbb{E}\left[\mathrm{e}^{-wY\frac{m}{n}}\right]\right)^m \qquad (7)$$

and $Y \in \mathrm{Exp}(1)$. So sum S behaves like the sum of m indep. $\mathrm{Exp}(n/m)$.

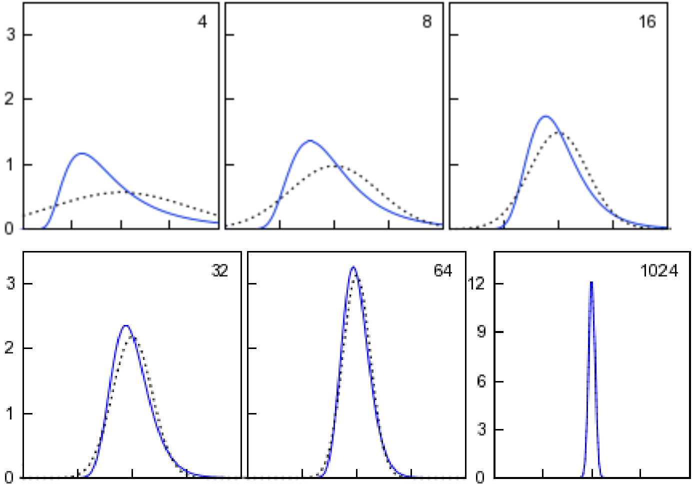
Thus, the rescaled/inverted $\frac{1}{n/m \cdot S}$ has induced density:

$$\overline{w}_m(u) = \mathrm{e}^{-1/u}\frac{u^{-m-1}}{(m-1)!}. \qquad (8)$$

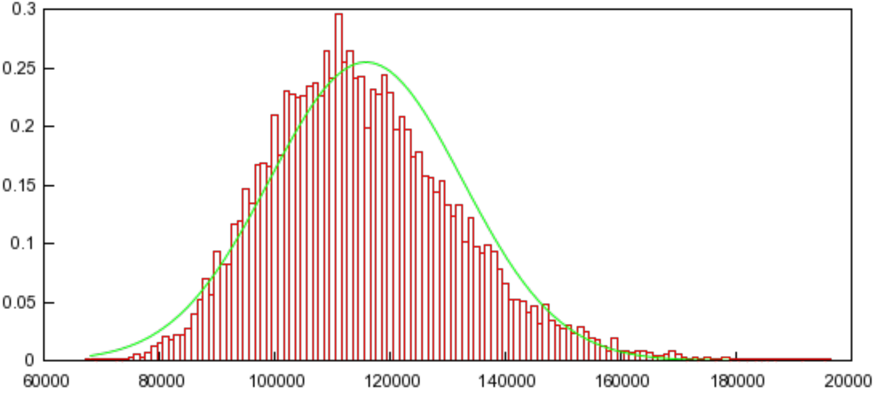
Theorem 3: For a fixed $m > 1$, as n tends to infinity, the estimator $\mathcal{Z}^\star$ satisfies

$$\lim_{n\to\infty} \mathbb{P}_n\left[\frac{\mathcal{Z}^\star}{n} \leqslant y\right] = \int_0^{y/(m-1)} \mathrm{e}^{-1/u}\frac{u^{-m-1}}{(m-1)!}\mathrm{d}u.$$

LIMIT DISTRIBUTION versus GAUSSIAN for $m = 4..1024$



OBSERVED estimations for $m = 50$

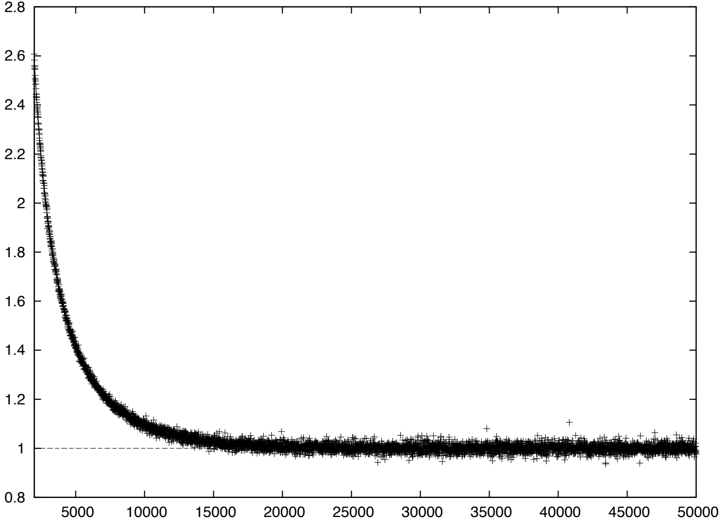


Cardinality estimation for an English dictionary, $n = 115\,794$ and $m = 50$.

Gaussian with $\mu = 115\,794$ and $\sigma = 1/\sqrt{48}$.

5. NON-ASYMPTOTICAL corrections

Plot of $\mathcal{Z}^\star/n$ relative to $n = 2000..50000$ (with $m = 4096$)



A) POISSON model (SOME urns are empty)

Pre-asymptotic calculations: n relatively small compared to m

Empty urns **bias** the average $\longrightarrow$ **keep track** + **ignore empty urns**.

Poisson approx. of urn allocation (i.e.: $N_j \in \mathrm{Poi}(\lambda)$).

$$\mathbb{P}[M_j \in [x, x + \mathrm{d}x]] = \lambda \mathrm{e}^{-\lambda x} \mathrm{d}x + \mathrm{e}^{-\lambda} \mathbf{1}_{\{x=1\}} \tag{9}$$

Let $k = \#\{\text{non-empty urns}\}$:

$$\mathbb{E}\left[\frac{1}{M_1 + \ldots + M_m}\right] = \sum_{k=0}^{m} \int_0^{\infty} \binom{n}{k} \left(\lambda \cdot \frac{1 - \mathrm{e}^{-a-\lambda}}{a + \lambda}\right)^k \left(\mathrm{e}^{-a-\lambda}\right)^{m-k} \mathrm{d}a$$

after calculations + Laplace, yields

Theorem 4: let $n/m = \lambda$ be such that $0 < \lambda < C < \infty$, then

$$\mathbb{E}_n[\mathcal{Z}^\star] \sim \frac{n}{1 - \mathrm{e}^{-\lambda}}. \tag{10}$$

B) "Linear counting"[1] (TOO MANY urns are empty)

Non-asymptotic: shift in point-of-view

n balls are thrown into m urns.

Classical result: Let $W_k := \#\{\text{urns containing } k \text{ balls}\}$

$$\mathbb{E}[W_k] \sim m \cdot \left[\frac{\lambda^k}{k!}\exp(-\lambda)\right]$$

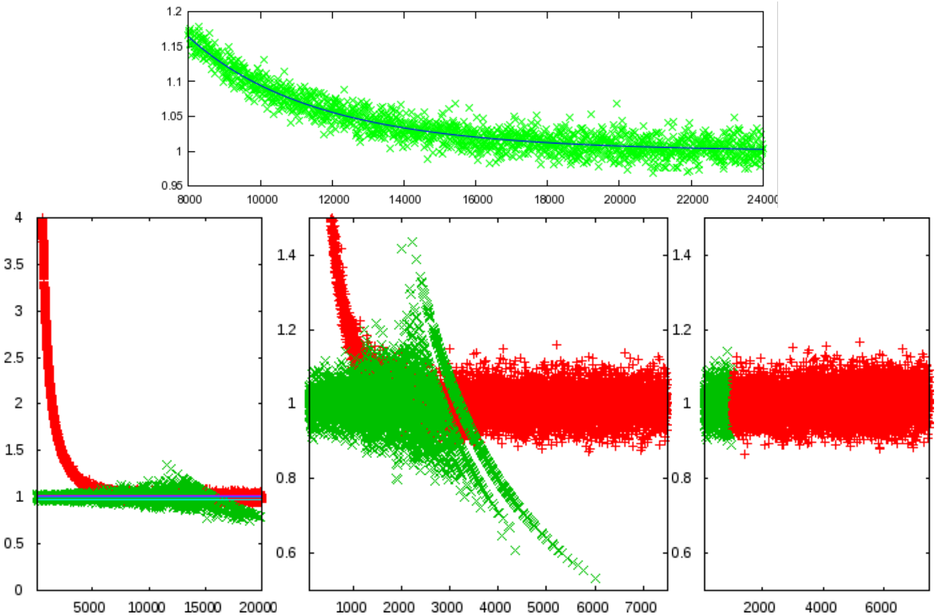
where $\lambda := n/m$ is constant, with $n \to \infty$ and $m \to \infty$.

Since $\mathbb{E}[W_0] \sim m \cdot \exp\left(-\frac{n}{m}\right)$ then, with $\widehat{w}_0$ observed empty urns,

$$n \approx -m \log\left(\frac{\widehat{w}_0}{m}\right) \tag{11}$$

[1] after Whang et al, "*A Linear-Time Probabilistic Counting Algorithm for Database Applications*," ACM Trans. on Database Systems, Vol. 15, No. 2, pp. 208-229, June 1990.

C) JOINING all regimes



linear counting

core algorithm

5. CONCLUSION

- a **complete** algorithm (large range of cardinalities)
- **optimal** within its class
- the **full** analysis with precise knowledge of the **output**

5. CONCLUSION

- a **complete** algorithm (large range of cardinalities)
- **optimal** within its class
- the **full** analysis with precise knowledge of the **output**

Future (immediate):

- **rate of convergence**
- attempt transposing the analysis to Hyperloglog
- plug-in the limit distribution analysis in simple algorithms which use cardinality estimation as a black box.