Beyond Distinct Counting: LogLog Composable Sketches of frequency statistics

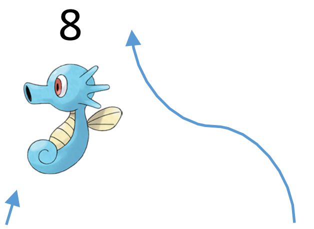
Edith Cohen

Google Research

Tel Aviv University

TOCA-SV 5/12/2017

Un-aggregated data



- Data element $e \in \mathrm{D}$ has key and value *(e.key,e.value)*

- Weight "frequency" of a key x: $w_x = \sum_{e \in \mathrm{D} | e.key = x} \mathrm{e.value}$

- Will also use:

Max $m_x = \max_{e \in D | e.key = x} \mathrm{e.value}$

"density" W of frequencies of keys)

Aggregated View

2×	7
1×	11
1×	17

f-statistics: $f(W) = \sum_{x \in \mathrm{X}} f(w_x)$

f - statistics $f(W) = \sum_{x \in X} f(w_x)$

- Distinct $f(\mathrm{w}) = 1$ $(\mathrm{w} > 0)$ #of distinct keys
- Sum $f(\mathrm{w}) = \mathrm{w}$
- Frequency moments $f(\mathrm{w}) = \mathrm{w}^p$
- Cap: $f(\mathrm{w}) = \min(\mathrm{T}, \mathrm{w})$
- Complement Laplace transform: $f(\mathrm{w}) = 1 - \mathrm{e}^{-\mathrm{wt}}$
- Other: $f(\mathrm{x}) = \log(1 + x)$ $f(x) = \min(x^{0.75}, T)$

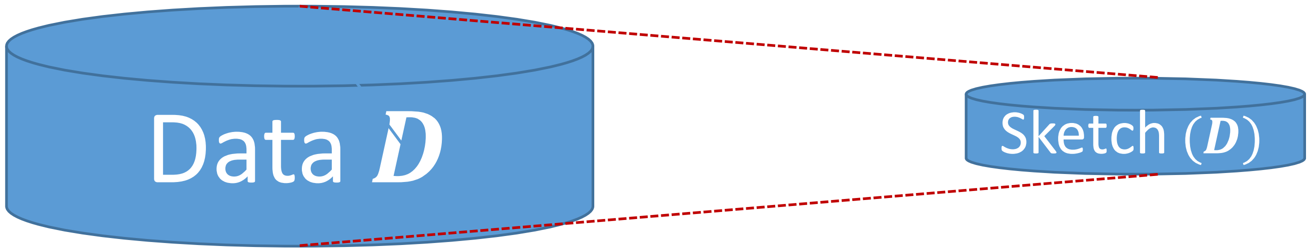
Applications: Search queries, online interactions, online transactions, word/term (co)-occurrences in corpus, network traffic

Issue: Aggregated view W is costly: Data movement, storage, computation

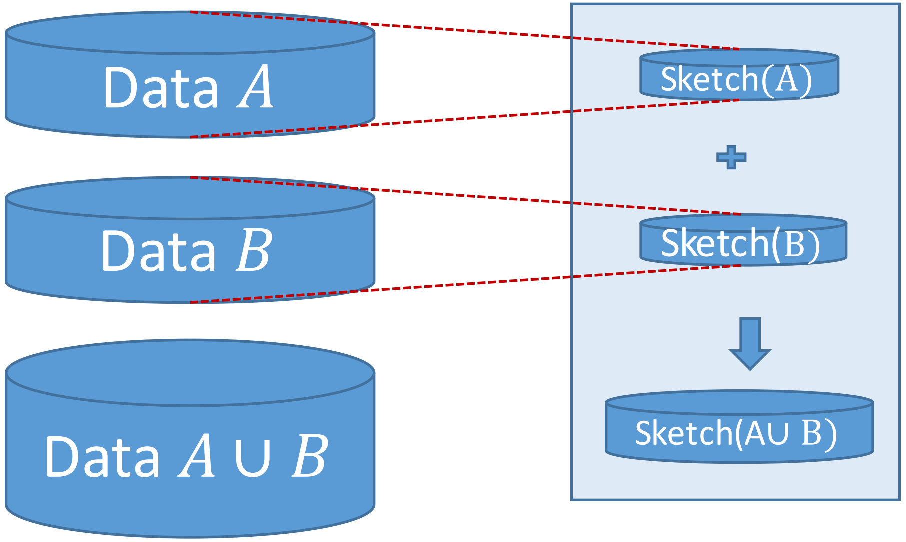
Use sketches!

Sketches

A sketch for f is a lossy summary of the data $\boldsymbol{D}$ from which we can approximate (estimate) $f(\boldsymbol{D}) = f(\boldsymbol{W})$



Q: $f(\boldsymbol{W})$ **?** → estimator $\hat{f}(\boldsymbol{S})$

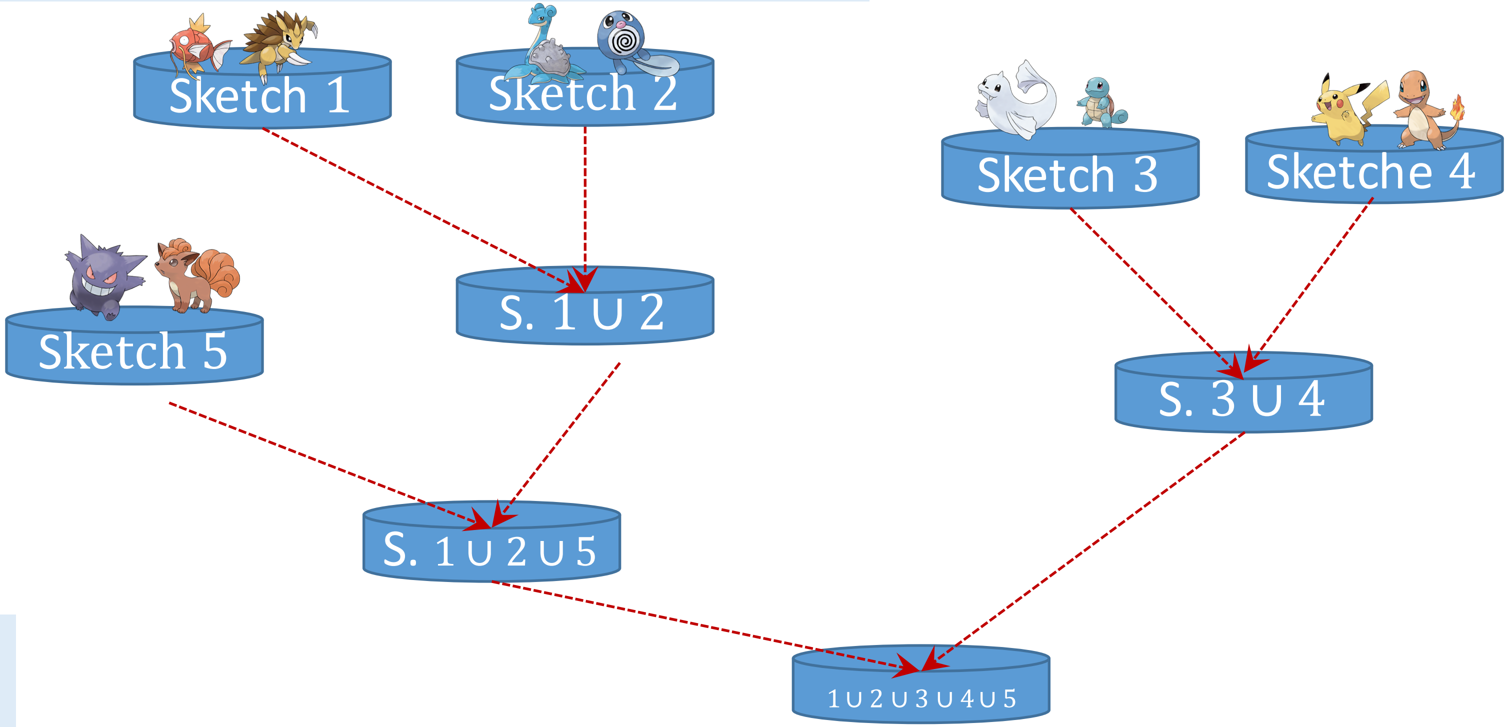


<u>Sketch structure design goals:</u>

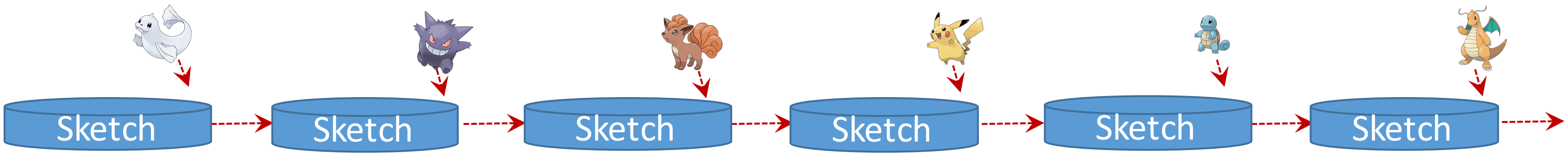
- Optimize sketch-size vs. estimate quality
  - Can we get size $O(\epsilon^{-2} + \log\log n)$?
- Composable/mergeable

Why composable is useful ?

Distributed data/parallelize computation



Streamed data



Approximate statistics via small sketches

- Data element $e \in D$ has key and value *(e.key,e.value)*
- Weight of key x is the Sum of its element values: $w_x = \sum_{e \in D | e.key = x} e.value$
- f-statistics: $f(D) = f(W) = \sum_{x \in X} f(w_x)$

Quality: Coefficient of variation $\frac{\sigma}{\mu} = \epsilon$, concentration

- ✔ Distinct $f(w) = 1$ $(x > 0)$: [Flajolet Martin '85, Flajolet et al '07] $O(\epsilon^{-2} + \log \log n)$
- ✔ Sum $f(w) = w$: [Morris '77] $O(\epsilon^{-2} + \log \log n)$
- Frequency moments $f(w) = w^p$: [Alon Matias Szegedy '99, Indyk '01] $O(\epsilon^{-2} \log^2 n)$
- **?** Capping $f(w) = \min(T, w)$ [C' 15] (via sampling) $O(\epsilon^{-2} \log n)$
- Others: "universal" sketches [Braverman Ostrovsky '10] Polynomial$(\epsilon^{-1}, \log n)$

Sum: $\sum_{x \in X} f(w_x)$

$\log(n)$: A single register of size to keep the sum. Clearly composable

$O(\epsilon^{-2} + \log\log n)$: [Morris 1977] +[Flajolet 1985] Composable version [C' 15]:

Maintain the "exponent" t, initialized $t \leftarrow 0$

- Estimate: return $(1+\epsilon)^t - 1$
- Add Y:
  - Increase t by maximum amount so that estimate increase by $Z \leq Y$
  - Let $\Delta = \mathrm{Y} - \mathrm{Z}$
  - Increment t with probability $\frac{\Delta\epsilon}{(1+\epsilon)^t}$
- Merge $t_2 \leq t_1$:

  same as Add $(1+\epsilon)^{t_2} - 1$ to counter t_1

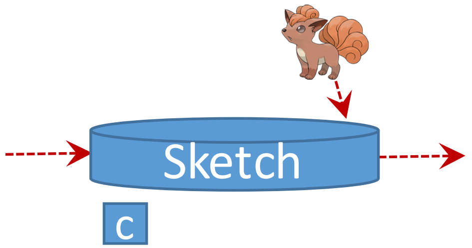
Distinct count sketches

HyperLogLog [Flajolet et al 2007]

- Optimal size $O(\epsilon^{-2} + \log\log n)$ for CV $\frac{\sigma}{\mu} = \epsilon$; n distinct keys
- **Idea:** store $k = \epsilon^{-2}$ exponents of hashes. Exponents value concentrated so store one and k offsets.

HIP estimators [Cohen '14, Ting '15]: halve the variance to $\frac{1}{2k}$!

- **Idea:** track an estimated count c with sketch structure. When structure is modified, add inverse modification probability to c.



Approx. count

If structure S is modified

$c \mathrel{+}= 1/p(\text{modified})$

Distinct count sketches

*Simplified version [C' 94]

- Initialize: $k = \epsilon^{-2}$ registers $c_1, \ldots, c_k, \leftarrow \infty$;
  - Hash functions $H_i(x) \sim Exp[1]$
- Process element $e.key$:
  - For i ∈ $[k]$: $c_i \leftarrow \min(c_i, H_i(e.key))$
- Estimate: $\frac{k-1}{\sum_i c_i}$

Analysis:

- $c_i \sim$ minimum over active keys of independent $EXP[1] \Longrightarrow c_i \sim EXP[\mathrm{Distinct}(D)]$
- Parameter estimation problem

Reduce size: keep exponents only of c_i , one exponent and ϵ^{-2} constant-size offsets

Composability: minimum is composable

MaxDistinct sketches

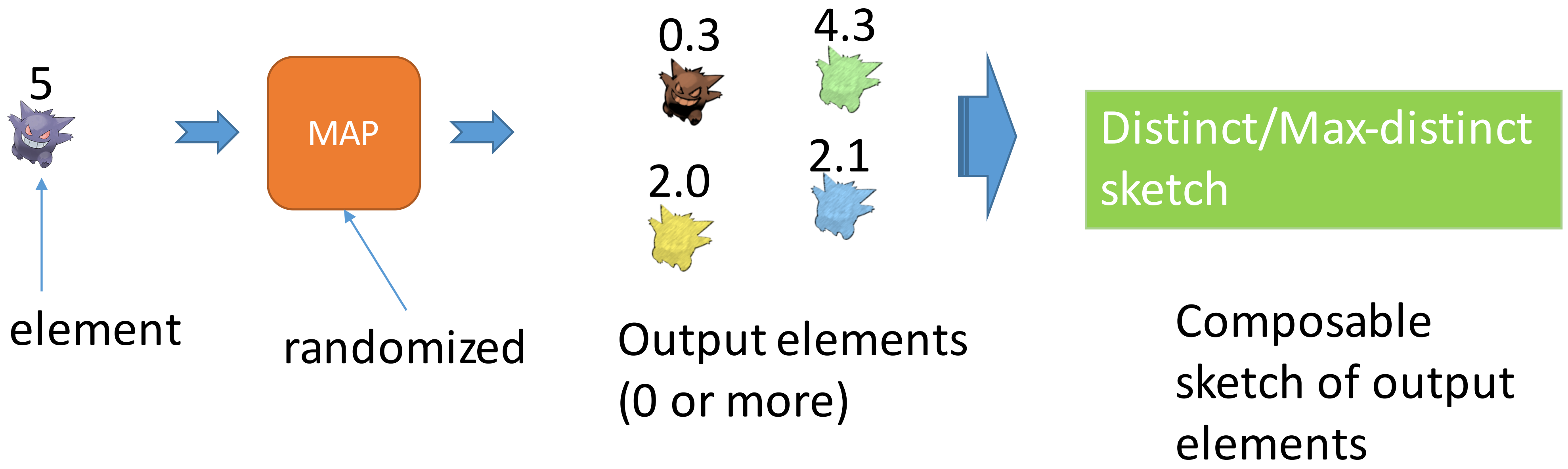
- Max value of an element with key x: $m_x = \max_{e \in D | e.key = x} \text{e.value}$
- MaxDistinct$(D) = \sum_x m_x$

- Initialize:
  - $k = \epsilon^{-2}$ registers $c_1, \ldots, c_k, \leftarrow \infty$
  - Hash functions $H_i(x) \sim Exp[1]$
- Process element $(e.key, e.value)$:
  - For i ∈ $[k]$: $c_i \leftarrow \min(c_i, \frac{H_i(e.key)}{e.value})$
- Estimate: $\frac{k-1}{\sum_i c_i}$

Analysis:

- For each key x, the minimum over elements of $\frac{H_i(e.key)}{e.value} \sim EXP[m_x]$
- $c_i \sim$ minimum over keys x of independent $EXP[m_x] \Longrightarrow$
  $c_i \sim EXP[MaxDistinct(D)]$

Element processing framework



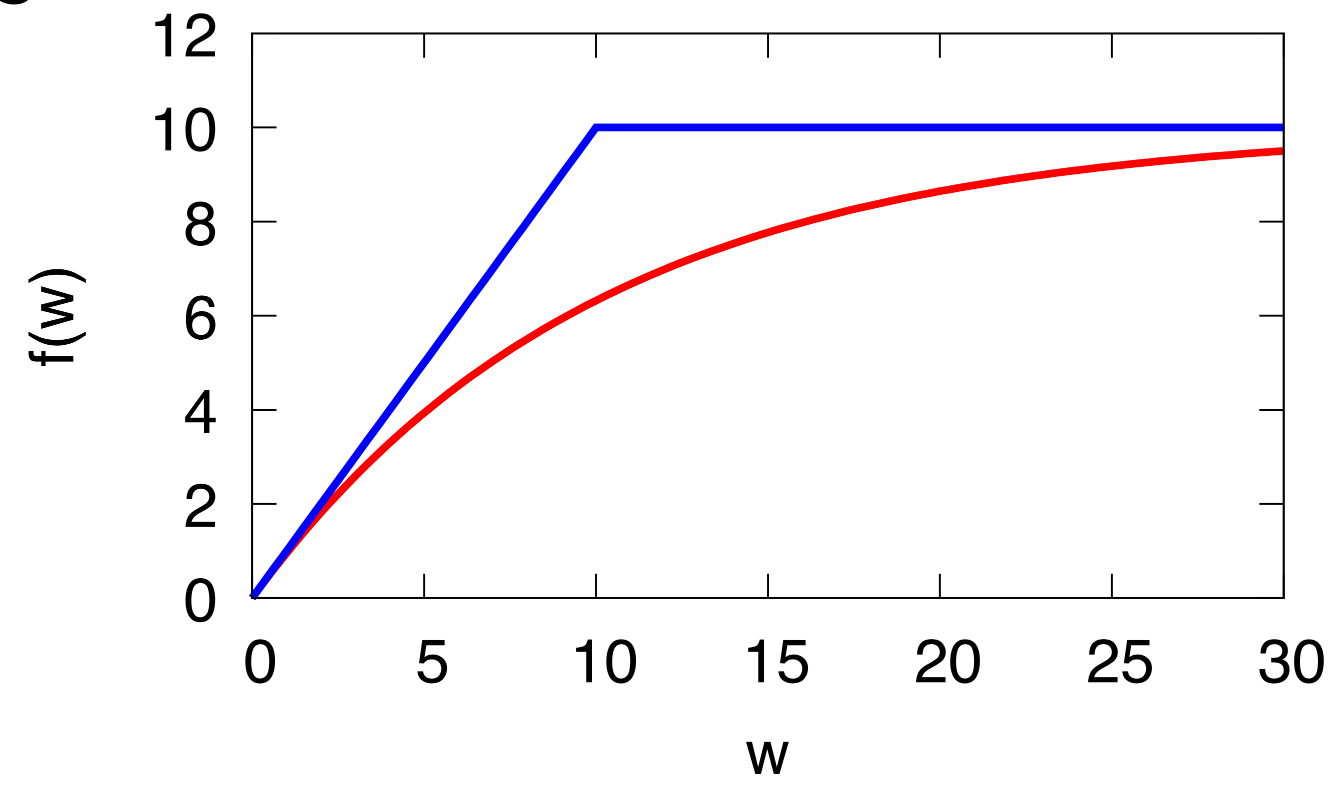
Goal: $\mathrm{E}[\ \mathrm{MaxDistinct}(\bigcup_{e\in D} \mathbf{MAP}(e))] = f(W)$, + concentration

Q: For which f we can do this? How? (specify MAP)

(Soft) cap functions

$$\widetilde{\text{cap}}_T(w) = T\left(1 - e^{-\frac{w}{T}}\right)$$

$$\text{cap}_T(w) = \min(T, w)$$



$W =$ 

(aggregated D)

$$\text{cap}_3(W) = \sum_{x \in W} \min(3, w_x) = 10$$

$$\widetilde{\text{cap}}_3(W) = 3 \sum_{x \in W} (1 - e^{-\frac{w_x}{3}}) \approx 8.38$$

Warm Up: Sketching $\widetilde{\text{cap}}_T$-statistics

we work with $f_t(w) = 1 - e^{-wt}$

!! The statistics is the Laplace[c] (complement-Laplace) transform of W (density function of frequencies)

$$L^c[W](t) = \sum_x (1 - e^{-w_x t})$$

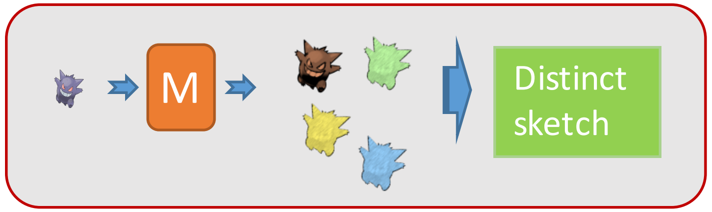
$\widetilde{\text{cap}}_T$-statistics is $T\times$ Laplace[c] transform at point $t = \frac{1}{T}$

since $\widetilde{\text{cap}}_T(w) = T f_{1/T}(w)$

$$\widetilde{\text{cap}}_T(W) = \sum_x \widetilde{\text{cap}}_T(w) = T \sum_x f_{\frac{1}{T}}(w_x) = T L^c[W](1/T)$$

Sketching Laplace^c transform of W at point t

$$f_t(w) = 1 - e^{-wt}$$



Element Map

Input: $e = (e.key, e.value)$

For $i = 1, \ldots, r$

- $y_i \sim \mathrm{EXP}[e.value]$
- **If** $y_i \le t$: **output** e.key#i

Output: (approximate) number of distinct output keys

! Output sketch size is barely affected by r, only element processing

Claim: (correctness)

$$\frac{1}{r}\mathrm{E}\left[\text{Distinct}\left(\bigcup_{e\in D}\mathbf{MAP}(e)\right)\right] = \sum_x 1 - e^{-w_x t} = L^c[W](t)$$

Element Map

Input: $e = (e.key, e.value)$

For $i = 1, \ldots, r$

- $y_i \sim \mathrm{EXP}[e.value]$
- **If** $y_i \le t$: **output** e.key#i

!! Each input key x and $i \in [r]$ have a unique potential output key $\boldsymbol{x\#i}$

We compute the probability that the output key $\boldsymbol{x\#i}$ is generated:

$\Leftrightarrow$ $\mathrm{y_i} \le t$ for at least one element $e \in D$ with $e.\mathrm{key} = x$

$$\Leftrightarrow \min_{\mathrm{e|e.key=x}} \mathrm{EXP}[e.value] \le t$$

$$\Leftrightarrow \mathrm{EXP}[w_x] \le t = 1 - \mathrm{e}^{-\mathrm{w_x t}}$$

Sum over all x, i (Poisson rvs) to establish claim

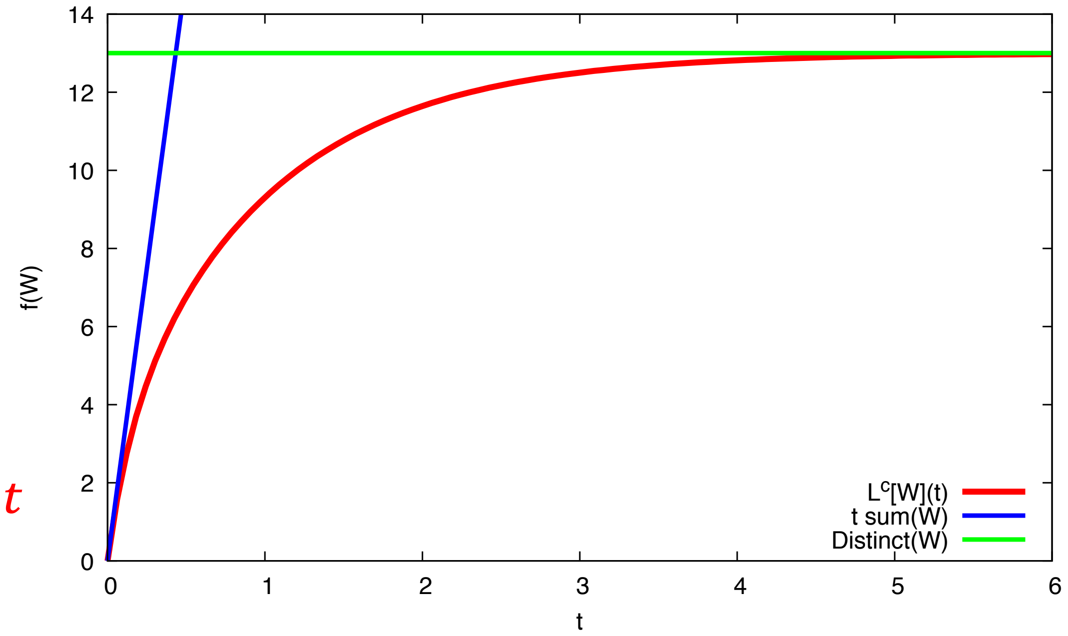
Subtlety:

? We need "enough" $\geq \epsilon^{-2}$ distinct output keys for low error.
We set $r = O(\epsilon^{-2})$, but still need to address small t …

- $\frac{1}{r}\mathrm{E}[\ \mathrm{Distinct}(\cup_{e\in D}\ \mathbf{MAP}(e))] = \sum_x 1 - e^{-w_x t} = \mathrm{L^c[W](t)}$

"density" W of frequencies:
$10\times\ \ w_x = 1$
$2\times\ \ w_x = 5$
$1\times\ \ w_x = 10$

- $Sum(W) = 30$
- $Distinct(W)=13$
- $L^c[W](t) = 13 - 10e^{-t} - 2e^{-5t} - e^{-10t}$



At the regime with low distinct count we can use $L^c[W](t) \approx tSum(W)$

$f(w)$ in the nonnegative span of $g_t(w) = 1 - e^{-wt}$

- Any function $f(w)$ that can be expressed $a(t) \geq 0$ as:

$$f(w) = \int_0^\infty a(t)(1 - e^{-wt})dt = L^c[a](w)$$

We get $a(t) = L^{c^{-1}}[f(w)](t) = \frac{1}{t} L^{-1}[\frac{\partial f(w)}{\partial w}](t)$

Span includes all concave sublinear f without “sharp” corners:
low frequency moments $p \leq 1$; logarithms; soft capping functions …

$f(w)$	$T(1 - e^{-\frac{w}{T}})$	$\sqrt{w}$	$\log(1 + w)$
$a(t)$	$T\delta(t - \frac{1}{T})$	$\frac{1}{2\sqrt{\pi}} \mathrm{t}^{-1.5}$	$\frac{e^{-t}}{t}$

Sketching statistics for f in the nonnegative span...

$$f(w) = \int_0^\infty a(t)(1 - e^{-wt})dt = L^c[a](w)$$

$$a(t) \geq 0$$

$$f(W) = \int_0^\infty f(w)W(w)dw = \int_0^\infty W(w) \int_0^\infty a(t)(1 - e^{-wt})dt\, dw =$$

$$= \int_0^\infty a(t) \int_0^\infty W(w)(1 - e^{-wt})\, dw\, dt$$

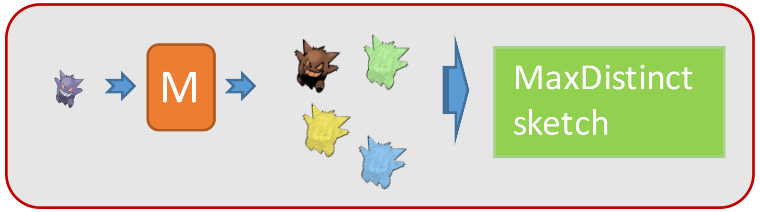
$$= \int_0^\infty a(t)\, L^c[W](t) dt$$

statistics $f(W)$ expressed in terms of the L^c transform

$\Rightarrow$ Can sketch $L^c[W](t)$ at many points t. But we will see a better way...

Sketching $f(W) = \int_0^\infty a(t)\, L^c[W](t)\,dt$

- Idea: Modify the element map for point t to work with weighted combination of t values



point t

Input: $e = (e.key, e.value)$

For $i = 1, \ldots, r$

- $y_i \sim \mathrm{EXP}[e.value]$
- **If** $y_i \leq t$: **output** e.key#i

Distinct count sketch

combination a(t) *slightly simplified

Input: $e = (e.key, e.value)$

For $i = 1, \ldots, r$

- $y_i \sim \mathrm{EXP}[e.value]$
- **output** (e.key#i, $\int_{y_i}^{\infty} a(t)\, dt$)

MaxDistinct sketch

Extension: Multi-objective sketch of the span

- Idea: If we can sketch all t values together, we can use the sketch for all statistics in the span
- log factor increase in sketch size (analysis of all-distance sketches [C' 94,15])

point t

Input: $e = (e.key, e.value)$

For $i = 1, \dots, r$

- $y_i \sim \text{EXP}[e.value]$
- **If** $y_i \le t$: **output** e.key#i

Distinct count sketch

All t

Input: $e = (e.key, e.value)$

For $i = 1, \dots, r$

- $y_i \sim \text{EXP}[e.value]$
- **output** (e.key#i,y_i)

"All-threshold" sketch

All-threshold Distinct sketches

Input elements $(e.key, e.value)$

Sketch allows us to approximate for any t : $\text{Distinct}\{e.key \mid e.value \leq t\}$

Distinct counting sketch

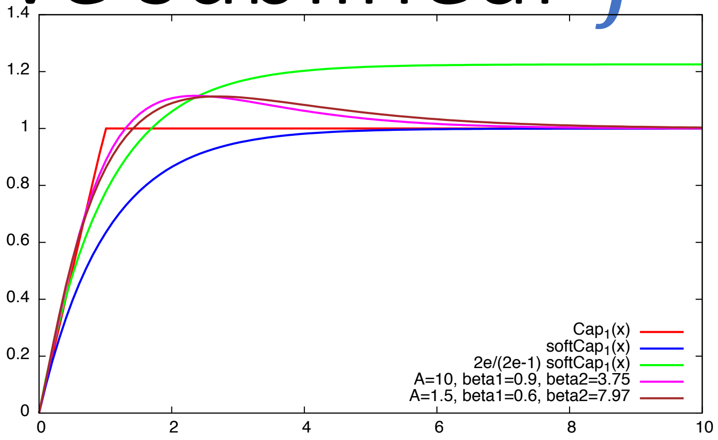
- Initialize: $k = \epsilon^{-2}$ registers $c_1, \ldots, c_k, \leftarrow \infty$;
  - Hash functions $H_i(x) \sim Exp[1]$
- Process element $e.key$:
  - For i ∈ $[k]$: $c_i \leftarrow \min(c_i, H_i(e.key))$
- Estimate: $\frac{k-1}{\sum_i c_i}$

All-threshold Distinct sketch

"Remember" for each register c_i all "breakpoints" where the minimum increases

- Logarithmic number of breakpoints [C' 94]
- Any "sample-based" distinct sketch can be similarly extended [C' 15]

Handling sharp corners: All concave sublinear f

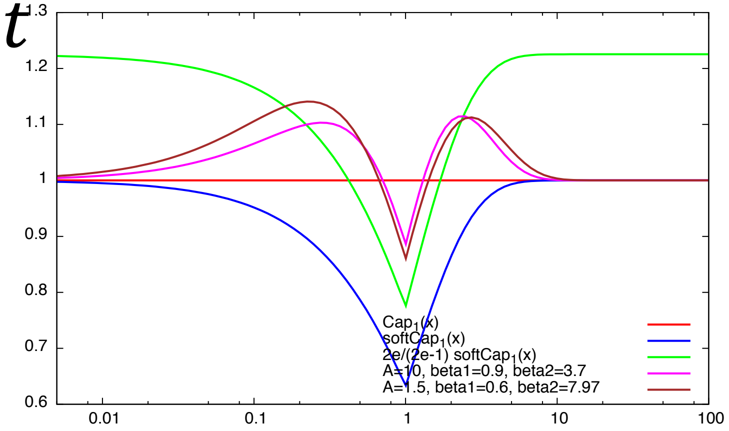


Reduces to sketching $\text{Cap}_1(w) = \min(1, w)$

- Soft cap gives $\left(1 - \frac{1}{e}\right) \times$ approximation

Better approximation:

- Use a signed inverse transform to approximate $\text{Cap}_1(w)$, controlling the $L_1(a(t)) = \int_0^\infty |a(t)|\, dt$
- Separate estimate the negative and positive components
- Grid search on 3 points $\Rightarrow$ We get 12%
- Open question to get ϵ



Conclusion

Summary:

- Simple, practical design of composable double-logarithmic size sketching for concave sublinear statistics
- Results novel theoretically even with $O(\epsilon^{-2} \log n)$ size

Open:

- Handling statistics with “sharp corners”.
- Loglog in the super-linear regime (second moment?)

Thank you !