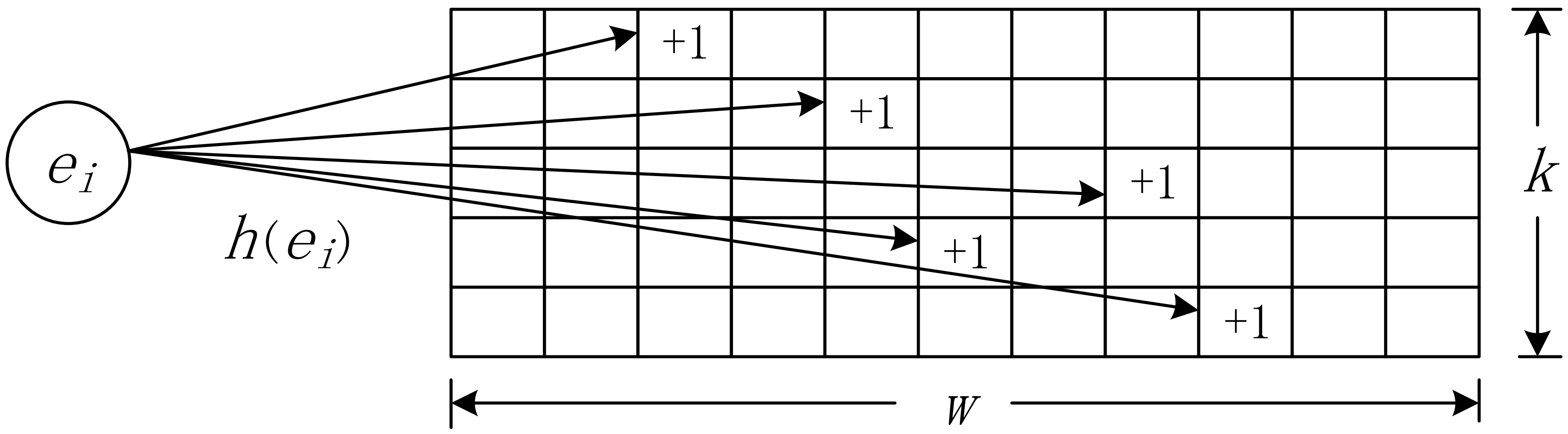
The Trapezoidal Sketch for Frequency Estimation in Network Flow

Ning Li
Harbin Institute of Technology

1. Motivation

- In many applications, the information of the streams needs to be recorded by the servers in real time.
- The accurate recording and estimation of the items' frequencies are always impractical or unnecessary.
- The sketch is one of the typical probabilistic data structures on estimating the frequency of items in data streams.

2. Problems

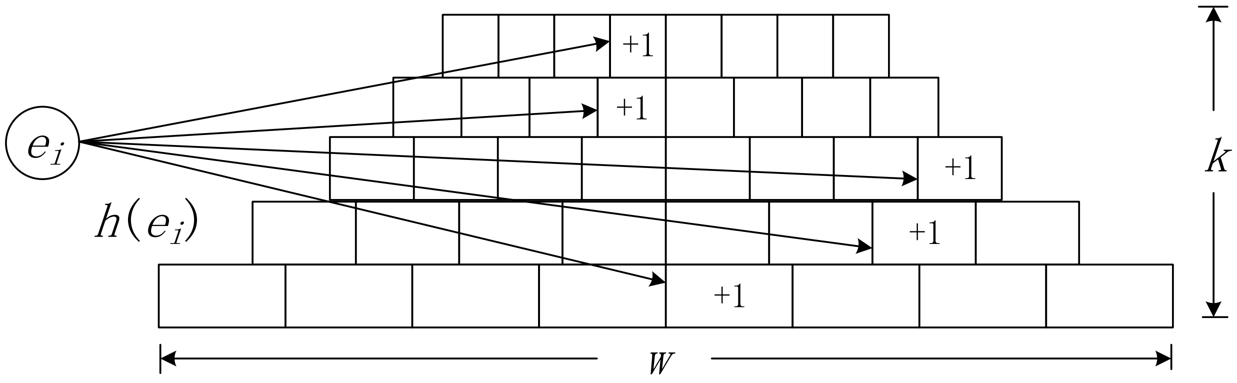


- The counter sizes in the *r*-sketch are the same, which is the inherent disadvantage of r-sketch and hard to be addressed
- The items’ frequencies are often highly skewed in real data streams

3. Proposed approaches

3.1. The basic t-sketch

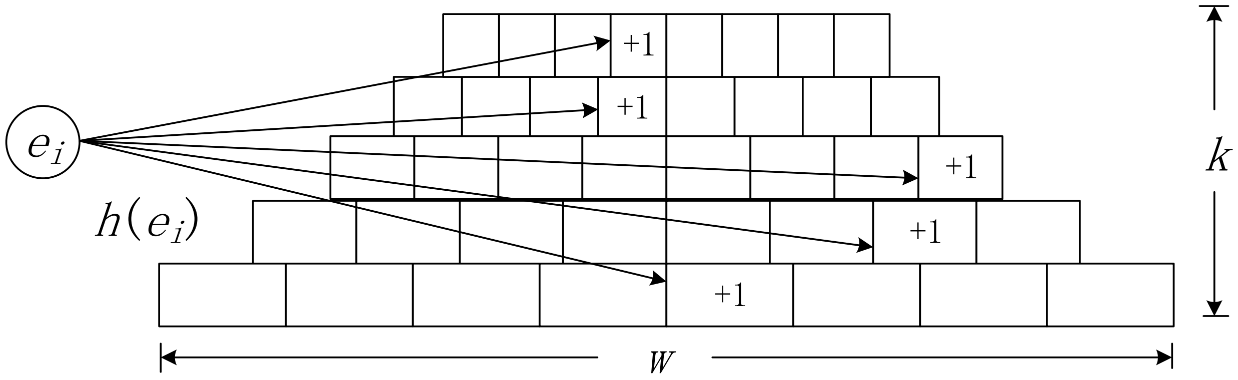
1. The sketch computes k hash functions $h_1(e)\%w,\ h_2(e)\%w,\ \cdots,\ h_k(e)\%w$ to determine k positions that the e_i is mapped to in the *t*-sketch;
2. Recording the values of these k counters;
3. Chosen the minimum value in these k counters as the estimated frequency of item e_i.



Property. In the *jth* layer of *t*-sketch, the probability that the estimation noise of item e_i cannot cause counter overflow is at least $1-\frac{\|f_{-e_i}\|_1}{w(S_j-f_{e_i})}$, i.e., $Pr\left[X_i \le S_j - f_{e_i}\right] \ge 1-\frac{\|f_{-e_i}\|_1}{w(S_j-f_{e_i})}$.

3. Proposed approaches

3.2. The space-saving t-sketch



Principle. The maximum counter size in space-saving t-sketch is the same as that in the r-sketch, i.e., B.

3. Proposed approaches

3.2. The space-saving t-sketch

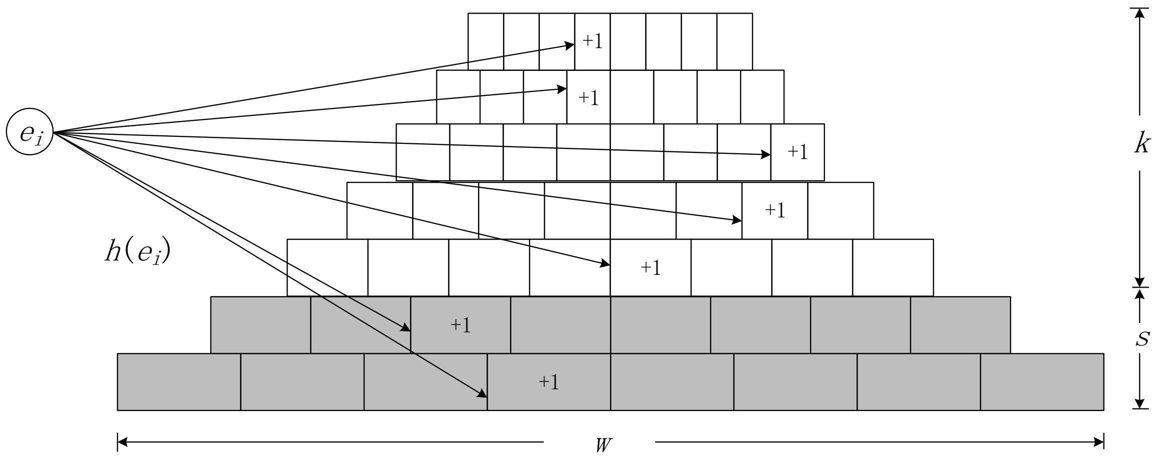
Property 1. The space-saving t-sketch can reduce the space usage and the reduction ratio is $\gamma = \frac{\log_2 d^{(k-1)}}{\log_2 B^2}$, where $d < B^{\frac{1}{k-1}}$.

Property 2. The probability that the estimated value is error in the space-saving t- sketch is $\rho_{sp-i} = \left(1 - \left(1 - \frac{1}{w}\right)^{n-1}\right)^{k-i}$, where i means that for the counters that element e_i mapped to, the counters in the first i layers are all overflow, i.e., from the first layer to the *ith* layer.

Property 3. In the space-saving t-sketch, the probability that the estimation error is smaller than $\beta\|f_{-e_i}\|_1$ is at least $1 - \frac{1}{(w\beta)^{k-i}}$, i.e., $Pr\left[\hat{f}_{e_i} - f_{e_i} \leq \beta\|f_{-e_i}\|_1\right] \geq 1 - \frac{1}{(w\beta)^{k-i}}$, where $\|f_{-e_i}\|_1$ is the frequencies of all the other items except e_i, $\hat{f}_{e_i}$ is the estimated frequency, f_{e_i} is the real frequency, and i indicates that the counter overflow occurs in *ith* layer.

3. Proposed approaches

3.3. The capacity-improvement t-sketch



Principle. The space usage in the capacity-improvement *t*-sketch is similar to that in the *r*-sketch

3. Proposed approaches

3.3. The capacity-improvement t-sketch

Property 1. The maximum counter size of the capacity-improvement t-sketch is $c = d^{\tilde{s}}B$, where $\tilde{s} = \lfloor s \rfloor$ and $s = \frac{\left[\left(\log_2 dB^2\right)^2 + \log_2 d^{4k(k-1)}\right]^{\frac{1}{2}} - \log_2 dB^2}{\log_2 d^2}$.

Three principles to decide the values of d and k.

1. If $\lfloor s_k^* \rfloor > \lfloor s_d^* \rfloor$, then $k^* = \log_2 B$ and $d^* = 2$.
2. If $\lfloor s_k^* \rfloor < \lfloor s_d^* \rfloor$, then $d = d^*$ and $k = \left\lfloor \frac{\log_2 B}{\log_2 d^*} + 1 \right\rfloor$.
3. When $\lfloor s_k^* \rfloor = \lfloor s_d^* \rfloor$, if $s_k^* - \lfloor s_k^* \rfloor > s_d^* - \lfloor s_d^* \rfloor$, then $k^* = \log_2 B$ and $d^* = 2$; otherwise, if $s_k^* - \lfloor s_k^* \rfloor < s_d^* - \lfloor s_d^* \rfloor$, $d = d^*$ and $k = \left\lfloor \frac{\log_2 B}{\log_2 d^*} + 1 \right\rfloor$.

3. Proposed approaches

3.3. The capacity-improvement t-sketch

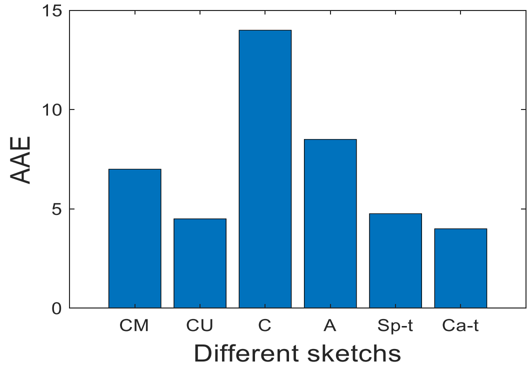
Property 2. The capacity of capacity-improvement *t*-sketch is $d^{\tilde{s}}$ times larger than the *r*-sketch.

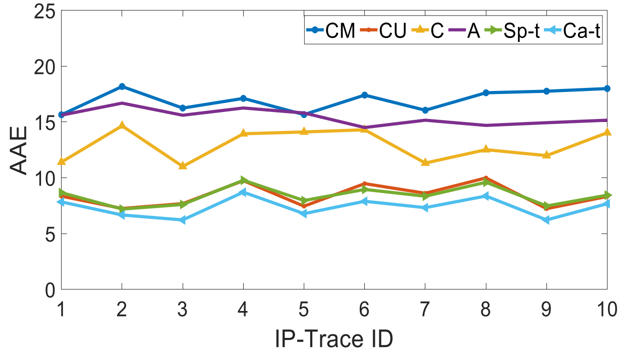
Property 3. For the capacity-improvement *t*-sketch, the saved space is $\theta_i = \log_2\left[B^{w(s_i^* - \lfloor s_i^* \rfloor)} \cdot d^{w(s_i^* - \lfloor s_i^* \rfloor)(s_i^* + 1)}\right]$, where $i = \{1,2\}$ represents principle_1 and principle_2.

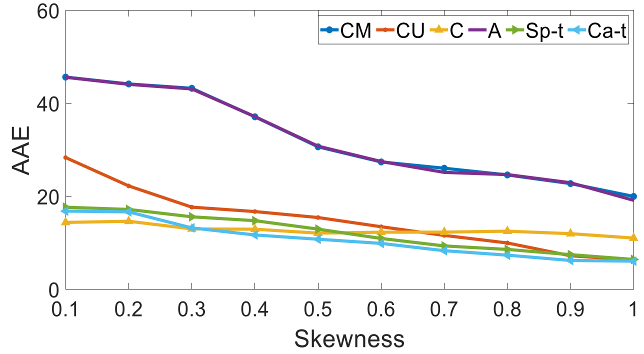
Error Probability. The error probability of the capacity-improvement *t*-sketch can be calculated as: $\rho_i = \left(1 - \left(1 - \frac{1}{w}\right)^{n-1}\right)^{k+\tilde{s}-i}$

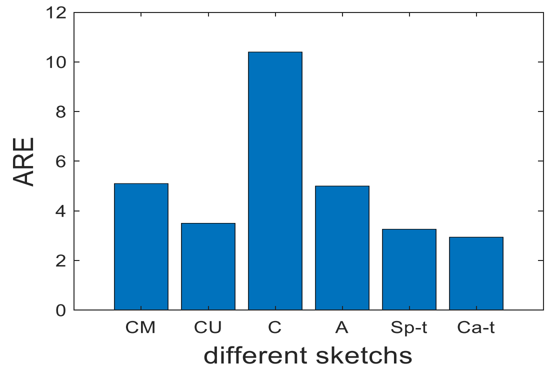
Estimation Error Boundary. The estimation error boundary of the capacity-improvement *t*-sketch can be calculated as: $Pr\left[\hat{f}_{e_i} - f_{e_i} \leq \beta \left\| f_{-e_i} \right\|_1\right] \geq 1 - \frac{1}{(w\beta)^{k+\tilde{s}-i}}$

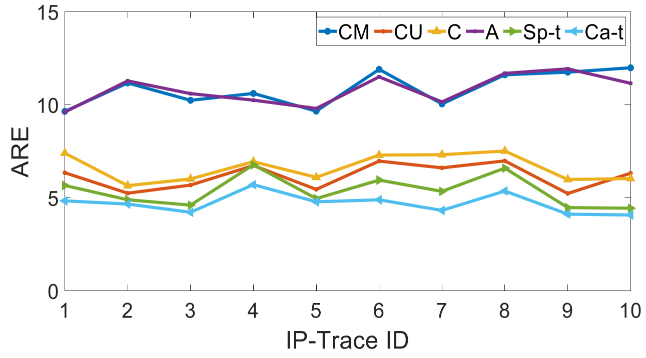
3. Simulation and Discussion

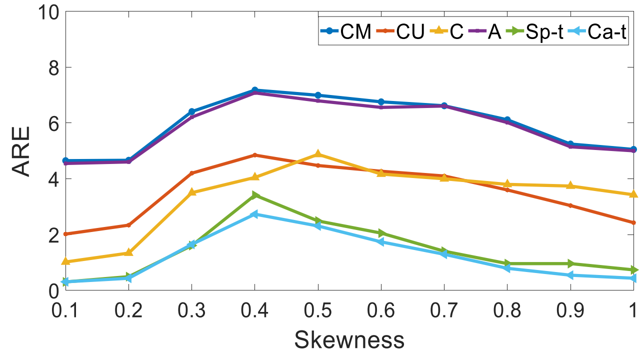












Thank you very much!