ECEN 689
Special Topics in Data Science for Communications Networks

Nick Duffield
Department of Electrical & Computer Engineering
Texas A&M University

Lecture 9
Count-Min Update

• When \((i(t), c(t))\) arrives:

\[ \text{add } c(t) \text{ to the element at column } h_j(i(t)) \text{ in each row } j \]

\[ K[ j, h_j(i(t)) ] + c(t) \text{ for } j = 1, 2, \ldots, d \]

\[ \begin{array}{cccccc}
\vdots & \vdots & + c(t) & \vdots & \vdots & \\
\vdots & \vdots & + c(t) & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \\
\vdots & \vdots & \vdots & \vdots & \vdots & \\
\end{array} \]


• Approximate \(a_i(t) = \sum_{s \leq t: i(s) = i} c(s)\) by \(\min_j K[ j, h_j(i) ]\) at time \(t\)
Dyadic Partitions and Range Queries

• Partition $n$ into **dyadic ranges** in $O(\log n)$ ways

  \{1,2,3,4,5,6,7,8\}

  \{1,2,3,4\} \quad \{5,6,7,8\}

  \{1,2\} \quad \{3,4\} \quad \{5,6\} \quad \{7,8\}

  \{1\} \quad \{2\} \quad \{3\} \quad \{4\} \quad \{5\} \quad \{6\} \quad \{7\} \quad \{8\}

• Any subinterval $[r,s]$ on $[1,n]$ can be expressed as a disjoint union of at most $2 \log n$ dyadic ranges $D$

• Maintain Count-Min sketch for each dyadic partition

• Estimate $a[r,s]$ by $a'[r,s] = \sum_{i=1}^{k} a'(D_i)$ where $a'(D_i)$ is Count-Min estimator for dyadic range $D_i$
Count-Min Sketch: Quantile Queries

• Consider \(a_i\) as weights of distribution over \(\{1, 2, \ldots, n\}\)

• \(\phi\)-quantile: \(j : a_1, \ldots, a_j\) have fraction \(\phi\) of total weight \(|a|_1\)

\[
\Sigma_{i=1, \ldots, j} a_i \leq \phi |a|_1 \leq \Sigma_{i=j+1, \ldots, nj} a_i
\]

• Example: median, with \(\phi = \frac{1}{2}\)

• Find approximate quantiles: tolerate error \(\varepsilon |a|_1\)

• Method: binary search on range sums
  – Exploit binary tree structure of dyadic sums
Count-Min Sketch Heavy Hitters

- **ϕ-heavy hitters**: elements \( i \) that have at least fraction \( \phi \) of total weight:

\[
a_i \geq \phi |a|_1
\]

- Find using search on dyadic tree:

\[
\{1,2,3,4,5,6,7,8\} \\
\{1,2,3,4\} \quad \{5,6,7,8\} \\
\{1,2\} \quad \{3,4\} \quad \{5,6\} \quad \{7,8\} \\
\{1\} \quad \{2\} \quad \{3\} \quad \{4\} \quad \{5\} \quad \{6\} \quad \{7\} \quad \{8\}
\]

- If \( a_i \) is a \( \phi \)-heavy hitter; all ancestors \( D \) of \( \{i\} \) have \( a[D] \geq \phi |a|_1 \)
Heavy Hitter Application

• Internet traffic
• Want to find dominant IP addresses
  – Originating or receiving large proportion of total traffic
  – Large website, DDOS victim
• Exploit natural tree structure of IP addresses based on prefixes
Counting distinct items in a stream

• Data stream \{a_1a_2a_3\ldots a_n\} of length n
• How many distinct values: m

• Applications
  – Number of distinct items as summary statistics of dataset
  – Useful for computing resource requirements
  – Compare values over different time windows, detect changes
Networking Applications

- Data stream = keys of packet stream
- #distinct items = #distinct keys = #flows
- Detecting distributed Denial of service attacks
  - Increase in number of SrcIP
- Detecting port scanning
  - Increase in number of DstPrt
Counting distinct elements in a stream

- Data stream \( \{a_1a_2a_3\ldots a_n\} \) of length \( n \)
- How many distinct values: \( m \)
- Exact approach:
  - Maintain hash table
  - Each arrival \( a \), store 1 in \( \text{hash}(a) \)
  - \( m = \#\text{entries in hash table} \)
- Storage cost: \( O(m) \)
- Approximate answer with less storage?
Probabilistic counting of distinct elements

- Data stream $S = \{a_1, a_2, a_3, \ldots, a_n\}$ of length $n$
- Hash $a \rightarrow h(a)$ uniform in $(0,1]$
- $m$ distinct values $\rightarrow m$ hashes IID uniform in $(0,1]$
- Maintain minimum value $H = \min_{a \in S} \{ h(a) \}$
- Streaming
  - Initialize $H = 1$
  - Foreach $a$; $H = \min\{H, h(a)\}$
- Estimate $m^* = 1/H$
Probabilistic counting

- By using hash $h(a)$ only distinct element are relevant
  - Multiple occurrences of same element get hashed to the same value
- More distinct values $\rightarrow$ Minimum tends to be closer to 0

- Consider $H_m = \min_{i=1,..,m} h_i$
  - Minimum of $m$ IID Unif[0,1] $h_i$

- CCDF $\Pr[H_m > x] = \Pr[\text{all } h_i > x] = (1-x)^m$

- PDF of $H_m$ is $m(1-x)^{m-1}$: $E[H_m] = 1/(m+1)$

- Using $m^* = 1/H$ is correct in some average sense
Probabilistic counting: bounds

- $m^* = \frac{1}{H_m}$ with $H_m = \min_{i=1,\ldots,m} h_i$

- Bound the probability that $m^*$ is lower than $m$ by a factor $k$

\[
\Pr[ m^* / m \leq 1/k ] = \Pr[\min_{i=1,\ldots,m} h_i \geq k/m ] \\
\leq \Pr[ h_i \geq k/m ]^m \\
= (1 - k/m)^m \\
\leq \exp(-k) < 1/k
\]
Probabilistic counting: bounds

- \( m^* = \frac{1}{H_m} \) with \( H_m = \min_{i=1,\ldots,m} h_i \)

- Bound the probability that \( m^* \) is higher than \( m \) by a factor \( k \)

\[
\Pr\left[ \frac{m^*}{m} \geq k \right] = \Pr\left[ \min_{i=1,\ldots,m} h_i \leq \frac{1}{(km)} \right] \\
\leq 1 - \Pr\left[ \min_{i=1,\ldots,m} h_i > \frac{1}{(km)} \right] \\
= 1 - \Pr[h_i > \frac{1}{(km)}]^m \\
= 1 - (1 - \frac{1}{(km)})^m \\
\leq \frac{1}{k}
\]

- Summary: \( \Pr[m^* \text{ over or under by factor } k] \) (each) \( \leq \frac{1}{k} \)
Discrete Probabilistic Counting in Practice

- Hash function $h$ into $[0,2^w -1]$, i.e. $w$ bit binary numbers
- For each $a$ in $S$, let $z(a) = \text{number of leading 0's of } h(a)$
- Define $Z = \max_a z(a)$
- Estimate $m^* = 2^Z$

Diagram:

- $h(a) = 0001...$  $z(a) = 3$
- $0 \rightarrow 2^{w-1}$
Discrete Probabilistic Counting in Practice

- $h(a)$ uniform in $[0, 2^w - 1]$ → bits of $h(a)$ IID with $\Pr[1] = \Pr[0] = 1/2$
- $\Pr[z(a) \geq r] = \Pr[\text{first } r-1 \text{ bits are 0}] = 2^{-r}$

- Let $x_a(r)$ be the indicator of the event $\{z(a) \geq r\}$:
  $$x_a(r) = 1 \text{ if } z(a) \geq r, \ 0 \text{ otherwise}$$

- Will need:
  $$\mathbb{E}[x_a(r)] = \Pr[z(a) \geq r] = 2^{-r} \text{ and } \text{Var}(x_a(r)) = 2^{-r} (1 - 2^{-r})$$

- Define $X(r) = \sum_a x_a(r)$:
  
  
  #distinct elements that lie in leftmost $1/2^r$ of $[0, 2^w - 1]$
Discrete Probabilistic Counting: Lower Bound

- \( \mathbb{E}[x_a(r)] = \mathbb{P}[z(a) \geq r] = 2^{-r} \) and \( \text{Var}(x_a(r)) = 2^{-r}(1-2^{-r}) \)
- \( X(r) = \sum_a x_a(r) : \text{#distinct elements in leftmost} \ 1/2^r \text{ of } [0,2^w-1] \)
- For \( c > 1 \), \( m^* > c \ m \) if \( 2^{z(a)} > cm \) for some \( a \), i.e.,

\[
\text{if } X(r) \geq 1 \text{ for some } r \text{ such that } 2^r > cm
\]

\[
\mathbb{P}[X(r) \geq 1] \leq \mathbb{E}[X(r)] \quad \text{(Markov)}
\]

\[
= \sum_a \text{distinct} \ \mathbb{E}[x_a(r)]
\]

\[
= m / 2^r
\]

\[
< 1/c
\]

- Conclusion: \( \mathbb{P}[m^* > c \ m] < 1/c \)
Discrete Probabilistic Counting: Upper Bound

• $E[x_a(r)] = \Pr[z(a) \geq r] = 2^{-r}$ and $\Var(x_a(r)) = 2^{-r} (1 - 2^{-r})$

• $E[X(r)] = mE[x_a(r)] = m2^{-r}$; $\Var[X(r)] = m\Var[x_a(r)] \leq m2^{-r} = E[X(r)]$

• For $c > 1$, $m^* < m/c$ if $2^{z(a)} < m/c$ for all $a$, i.e.,

  if $X(r) = 0$ for some $r$ such that $2^r < m/c$

  \[
  \Pr[X(r) = 0] = \Pr[ |X(r) - E[X(r)]| \geq E[X(r)] ] \\
  \leq \frac{\Var(X(r))}{E[X(r)]^2} \quad \text{(Chebychev)} \\
  \leq \frac{1}{E[X(r)]} \\
  = \frac{2^r}{m} < \frac{1}{c}
  \]

• Conclusion: $\Pr[m^* < m/c] < 1/c$
Better estimates through combination

• Single $m^*$ is quite noisy
  – $m^*$ is always a power of 2
  – Markov inequality is quite weak in general
• Improvement
  – Combine multiple $m^*$ computed in parallel with different hash functions
• Use median?
  – Median of $s$ estimates $\{m^*_{1}, m^*_{2}, \ldots, m^*_{s}\}$ is still a power of 2
• Use mean?
  – $m^*$ does not have good averaging properties
  – Probability to double $m^*$ from some $2^{z(a)}$ is $\frac{1}{2}$
    • Get contributions to $E[m^*]$ out to $w$
    • If we had unlimited bits then $E[m^*]$ would be infinite
• Best of both worlds
  – Compute median of multiple averages
    • Averages are not restricted to powers of two, median omits large values
References
