Counting distinct items in a stream

• Data stream \( \{a_1, a_2, a_3, \ldots, a_n\} \) of length \( n \)
• Want to find \( m = \) number of distinct values of the \( a_i \)

• 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 hash(a)
  - m = #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 \( m \) by \( m^* = 1/H \)
Probabilistic counting

• By using hash $h(a)$ only distinct elements are relevant
  – Multiple occurrences of same element get hashed to the same value
• If 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

- \( m^* = 1 / H_m \) 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

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

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

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

- Summary $Pr[m^* \text{ over or under by factor } k] \leq 2/k$
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 $z(a) =$ number of leading 0's of $h(a)$
  - $z(a) \geq r$ means $h(a)$ lies in leftmost $1/2^r$ fraction of $[0, 2^w - 1]$
- Define $Z = \max_a z(a)$
- Estimate $m^* = 2^Z$

\[ h(a) = 00001... \quad z(a) = 4 \]
Discrete Probabilistic Counting in Practice

• $h(a)$ uniform in $[0, 2^w - 1]$ → bits of $h(a)$ IID with $Pr[1] = Pr[0] = 1/2$

• $Pr[z(a) \geq r] = Pr[\text{first } r \text{ bits are 0}] = 2^{-r}$

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

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

- \( E[x_a(r)] = \Pr[z(a) \geq r] = 2^{-r} \) and \( \text{Var}(x_a(r)) = 2^{-r}(1-2^{-r}) \)
- \( X(r) = \sum_{a \text{ distinct}} x_a(r) \):
  - \#distinct elements in leftmost \( 1/2^r \) of \([0,2^w-1]\)
- Recall estimate \( m^* = 2^{\max z(a)} \) for \( z(a) = \# \text{ leading 0's in } h(a) \)
- 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 > c \) \( m \)
  \[
  \Pr[X(r) \geq 1] \leq E[X(r)] \quad \text{(Markov)}
  \]
  \[
  = \sum_{a \text{ distinct}} E[x_a(r)]
  \]
  \[
  = m / 2^r
  \]
  \[
  < 1/c
  \]
- Conclusion: \( \Pr[m^* > c \) \( m] < 1/c \)
Discrete Probabilistic Counting: Upper Bound

- \( E[x_a(r)] = \Pr[z(a) \geq r] = 2^{-r} \) and \( \text{Var}(x_a(r)) = 2^{-r} (1-2^{-r}) \)
- \( E[X(r)] = m E[x_a(r)] = m 2^{-r}; \Var[X(r)] = m \text{Var}[x_a(r)] \leq m 2^{-r} = E[X(r)] \)

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

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

- 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