Lecture 4  1/27/22

Perfect Hashing

Goal: Dist. with $O(n)$ space
$O(1)$ query time deerm.

Contrast: last time $O(1)$ expected.

Question: what if build a hash table
with random h.f. with no collision.

$C_x = \# \{y \in S \mid y \neq x \text{ and } h(y) = h(x)\}$.

$C = \sum_{x \in S} C_x =$ # pairwise col. $x, y \in S$.

Suppose we want $C = 0 \Rightarrow$ all buckets
have size $O(1)$.

$\mathbb{E}[C^3]$ when we have $m$ buckets.

$h : U \rightarrow [m^3]$.

$\mathbb{E}[C^3] = \mathbb{E}\left[ \sum_{x \in S} C_x \right] = \sum_{x \in S} \mathbb{E}[C_x^3]$.

$\leq \sum_{x \in S} \frac{n}{m}$

$= \frac{n^2}{m}$.  

$\text{SCU}$

$n = 181$

$m = \frac{\text{table size}}{\text{size}}$
By Markov bound: \( C \leq 4 \ln \left( 1 + \frac{1}{4} \right) \) with prob \( \geq 1 - \frac{1}{4} \).

Set \( m = 8n^2 \Rightarrow C \leq \frac{1}{2} \) with prob \( \geq 1 - \frac{1}{4} \).

\( \Rightarrow C = 0 \)

**Algo:**

1. **repeat until success**
2. pick random \( h \in \mathcal{H} \), \( m = 8n^2 \)
3. compute \( C \)
4. if \( C > 1 \), repeat.
5. if \( C = 0 \), success.
6. build hash table \( H \) with last \( h \).

**Algo** has space \( O(m+n) = O(n^2) \)

q.t.: \( O(1) \) determin.

**Preprocessing:** proportional to \# repeats:

\[
\mathbb{E} [\text{\# repeats}] \leq 1 + \frac{3}{4} + 2 \cdot \frac{1}{4} + \frac{3}{4} + 3 \left( \frac{1}{4} \right)^2 + \ldots
\]

\[
= 1 + \frac{1}{4} + \left( \frac{1}{4} \right)^2 + \ldots
\]

\[
= \frac{1}{1 - \frac{1}{4}} = \frac{4}{3}.
\]

**Cor:** can obtain \( O(n^2) \) space, \( O(1) \) deter. q.t.

**Perfect Hashing**
Idea: use standard hashing, \( m = 2n \). + 2\textsuperscript{nd} level hash where we split non-empty buckets using quadratic-space hash.

\[ h \]

\[ m_i = 8n_i^2 \]

\[ n_i = \# \text{clw's going bucket } i \]

**Algo PH:**

Pick a random h.f. \( h : U \to 1m3, m = 2n \).

For \( i = 1 \ldots m \):

- build 2\textsuperscript{nd} level hash table of size
  
  \[ m_i = 8n_i^2 \] (using Cor)

  where
  
  \[ n_i = |S \cap h^{-1}(i)| = \# \text{clw's in bucket } i \]

  store \( S \cap h^{-1}(i) \) in the 2\textsuperscript{nd} level h.t.

**Query time:** look-up \( h(x) \), then look-up hash table corresponding to \( i = h(x) \).

**Obs:** query time is \( O(1) \) def.

**Question:** what about space?
Claim: in expectation, size of 2nd level hash tables is \( O(n) \).

\[ n_i + n_i(n_i - 1) = n_i^2 \]

pf: \( \sum_{i=1}^{m} n_i = \sum_{i=1}^{m} 8n_i^2 \).

\[
\begin{align*}
\mathbb{E} \left[ \sum_{i=1}^{m} n_i^2 \right] &= \mathbb{E} \left[ \sum_{i=1}^{m} n_i + n_i(n_i - 1) \right] \\
&= \mathbb{E} \left[ \sum_{i=1}^{m} n_i \right] + \mathbb{E} \left[ \sum_{i=1}^{m} n_i(n_i - 1) \right] \\
&= \mathbb{E} \sum_{i=1}^{m} n_i + \mathbb{E} \left[ \sum_{i=1}^{m} n_i(n_i - 1) \right] \\
&= n + \mathbb{E} [C] \\
&= n + n^2/m = n + \frac{n^2}{m} = O(n).
\end{align*}
\]

Conclusions: PH algo obtains \( O(1) \) def. q.t.

\( O(n) \) space \( \exp \)

\( O(n) \) preproc. \( \exp \)

Remark: can have \( O(n) \) def. space bound by repeating \( \Sigma \) lev hashing until \( \Sigma n_i^2 \leq O(n) \).

Remark: dynamic Dct.: can \( \omega \) obtain \( O(1) \) def. query/update time?
Stream & Sketching

Motivation:

Goal: store statistic on data, as streams to be used later.

Distinct element count.

Stream: elements from universe $\mathbb{U} = \{1, \ldots, n\}$.

E.g.: $\mathbb{U}$ = all possible IPs.

Problems: report how many different elements we have seen in the stream.

Stream length: $m$.

Goal: store as little info as possible.

Solution 1: store entire stream: $O(m)$ space.

Solution 2: store a table $T[1..n]$. 

\textit{OPEN.}
Can we obtain space $\ll \min \{n, m\}$?

No unless we allow approx + random.

**Algo:** approx + random. [Flajolet - Martin]

- pick a random hash func. $h: [n] \rightarrow [0, 13].$
- store a reg. $z$, initial $z = 1$.
- when see $i$ in the stream:
  $$z = \min \{z, h(i)\}.$$
- $\text{Est} = \frac{1}{z} - 1.$

**Note:** $z = \min h(i)$ hash function value $h(i)$ among i's in the stream.

**Claim:** $\mathbb{E}[z] = \frac{1}{d+1}$, $d = \# \text{distinct els'.}$

**Pf:** $z = \min$ of $d$ random #'s $a_i \in [0, 13].$

**Experiment:** pick $a \in [0, 13].$

$$\mathbb{E}\left[\frac{1}{z}\right] = \mathbb{E}\left[\frac{1}{\mathbb{E}[rac{1}{a}]}\right]$$
\[ P_1 \left( a \leq 2 \right) \]

\[ \Rightarrow 2) \quad \text{Prob. that } a \text{ is } \min \left\{ \alpha_1, \alpha_2, \ldots, \alpha_d, q \right\} = \frac{1}{d+1} \]

\[ \Rightarrow \mathbb{E} \left[ E2 \right] = \frac{1}{d+1} \cdot \mathbb{E} \left[ E \right]. \]

\[ \forall \frac{1}{2} \neq \forall \mathbb{E} \left[ E \right]. \]