**Announcements**

1. Homework 6 to be released Weds, due one week later (plus 2-day grace period).
2. Quiz 4 will be in class, Monday, May 9.

**Recap:** $h(x) = ax + b \pmod{M}$, $M \geq m$, prime.

$(a, b) \in [M]^2$ uniformly random.

- Storing a representation of $h$ requires only $2 \log_2(M)$ space to store coefficients $a$ and $b$.

1. $\forall x$, $h(x)$ is uniformly distributed in $[M]$.
2. $\forall x \neq y$ then $h(x), h(y)$ are independent random variables.

Let $X_{i,k} = \begin{cases} 1 & \text{if } h(a_i) \leq k \\ 0 & \text{if } h(a_i) > k \end{cases}$

$Y_k = \sum_{i=1}^{d} X_{i,k} = \# \text{ of distinct elements in stream that hash to } \{h, \ldots, h_k\}$.

We proved:

1. $E[X_{i,k}] = \frac{k}{M}$
2. $E[Y_k] = \frac{dk}{m}$
3. $\text{Var}[Y_k] < \frac{dk}{M}$. 
Algorithm for Distinct Elements

Sample random hash function \( h \) as above.

Let \( t = \left\lfloor \frac{2(1+e)}{e^2} \right\rfloor \).

Initialize \((Z_1, \ldots, Z_t) = 1^t\).

// \( Z_1, \ldots, Z_t \) will store the \( t \) smallest distinct values in the set \( \{ h(a_i) \mid i = 1, \ldots, n \} \) in increasing order.

for \( i = 1, \ldots, n \):
    observe token \( a_i \);
    compute \( z = h(a_i) \);
    if \( z < Z_t \):
        update \( Z_1, \ldots, Z_t \) to preserve the invariant in the comment above.
endfor

output \( \frac{tM}{Z_t} \).

Analysis of the algorithm

\( Z_t = k \iff Y_k = t \text{ and } Y_{k-1} < t \).

So, we resort to analyzing \( Y_k \) for various values of \( k \).
Two ways the algo could fail:
- Outputs answer \( < (1-e) \cdot d \).
  Since \( \text{answer} = \frac{t \cdot M}{\pi t} \) this
  corresponds to \( \pi t > \frac{t \cdot M}{(1-e) \cdot d} \)
  \[ \Rightarrow Y_k < t \quad \text{for} \quad k = \left\lfloor \frac{t \cdot M}{(1-e) \cdot d} \right\rfloor \]
- Outputs answer \( > (1+e) \cdot d \)
  corresponds to \( \pi t < \frac{t \cdot M}{(1+e) \cdot d} \)
  \[ \Rightarrow Y_k > t \quad \text{for} \quad k = \left\lfloor \frac{t \cdot M}{(1+e) \cdot d} \right\rfloor \]

\[ E[Y_k] = \frac{d \cdot t}{M}, \quad \text{Var}[Y_k] < \frac{d \cdot t}{M} \]

\[ P_r(Y_k > t) \leq \frac{d \cdot t / M}{(d \cdot t / M - t)^2} \]

\[ \frac{d \cdot t}{M} = \frac{d}{M} \int_{1/1+e}^{1} \frac{M}{d} \, d \frac{t}{1+e} \leq \frac{t}{1+e} \]

\[ t - \frac{d \cdot t}{M} \geq t - \frac{t}{1+e} = \frac{et}{1+e} . \]

\[ (t - \frac{d \cdot t}{M})^2 \geq \frac{e^2 t^2}{(1+e)^2} . \]

\[ t \leq \left[ \frac{2(1+e)}{e^2 \delta} \right] \left\lfloor \frac{1+e}{e^2 t} \right\rfloor \leq \frac{\delta}{2} . \]

Similar calculation using Chebyshev shows the probability of answer \( < (1-e) \cdot d \) is also \( \leq \frac{\delta}{2} \).
Sketching Model of Computation

Stream a1, ..., an
Tokens belonging to \([m]\)
Also has storage space \(s = O(\text{poly}(\log n, \log m))\).

After processing stream, the stored representation is used to answer queries from some set \(Q\) of potential queries.

\((\epsilon, \delta)\)-PAC property: \(\forall q \in Q\) \(\Pr[\text{answer is } \epsilon\text{-accurate}] \geq 1 - \delta\).

Weds lecture: Estimating Frequencies of Tokens.
Query \(q(x)\) for token \(x\) asks,
"How many times did \(x\) appear in the stream?"
\(\epsilon\)-accurate might mean, for example, answer to \(q(x)\) has additive error \(\leq \epsilon n\).