

Advanced Algorithms

南京大学

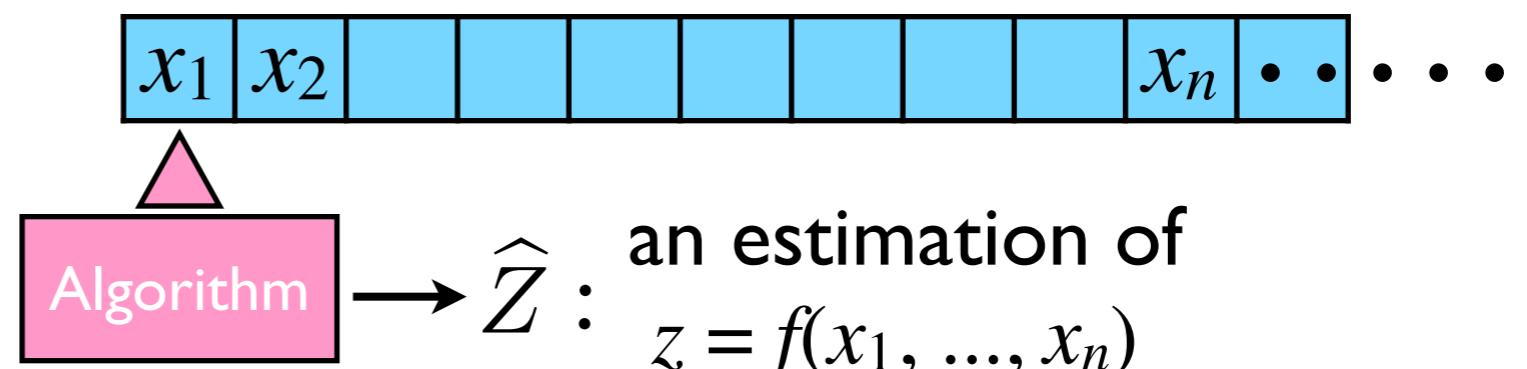
尹一通

Count Distinct Elements

Input: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \dots, x_n\}|$

- **data stream:** input comes one at a time
- **naive algorithm:** store everything with $O(n)$ space



- **(ϵ, δ)-estimator:** $\Pr \left[(1 - \epsilon)z \leq \hat{Z} \leq (1 + \epsilon)z \right] \geq 1 - \delta$

Using only memory equivalent to 5 lines of printed text, you can estimate with a typical accuracy of 5% and in a single pass the total vocabulary of Shakespeare. -----Flajolet

Input: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \dots, x_n\}|$

- **(ϵ, δ)-estimator:** $\Pr \left[(1 - \epsilon)z \leq \hat{Z} \leq (1 + \epsilon)z \right] \geq 1 - \delta$

uniform hash function $h: \Omega \rightarrow [0,1]$

$h(x_1), \dots, h(x_n)$: z uniform independent values in $[0,1]$

(partition $[0,1]$ into $z+1$ subintervals)

$$\mathbb{E} \left[\min_{1 \leq i \leq n} h(x_i) \right] = \mathbb{E}[\text{length of a subinterval}] = \frac{1}{z+1}$$

estimator: $\hat{Z} = \frac{1}{\min_i h(x_i)} - 1$? (by symmetry)

But $\text{Var}[\min_i h(x_i)]$ is too large!

(think of $z=1$)

Input: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \dots, x_n\}|$

- **(ϵ, δ)-estimator:** $\Pr \left[(1 - \epsilon)z \leq \hat{Z} \leq (1 + \epsilon)z \right] \geq 1 - \delta$

uniform independent hash functions:

$$h_1, h_2, \dots, h_k : \Omega \rightarrow [0,1] \quad Y_j = \min_{1 \leq i \leq n} h_j(x_i)$$

$$\text{average-min: } \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j$$

$$\text{Flajolet-Martin estimator: } \hat{Z} = \frac{1}{\bar{Y}} - 1$$

UHA: Uniform Hash Assumption

unbiased estimator: $\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$

- **Deviation:** $\Pr \left[\hat{Z} < (1 - \epsilon)z \text{ or } \hat{Z} > (1 + \epsilon)z \right] < ?$

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

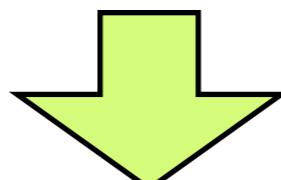
$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry}$$

$$\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$$

F-M estimator:

$$\text{let } \hat{Z} = \frac{1}{\bar{Y}} - 1$$

goal: $\Pr [\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z] < \delta$

 for $\epsilon \leq 1/2$

$$\left| \bar{Y} - \frac{1}{z+1} \right| > \frac{\epsilon/2}{z+1}$$

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

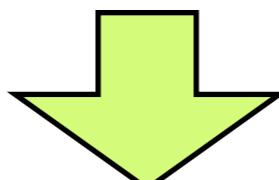
$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry}$$

$$\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$$

F-M estimator:

$$\text{let } \hat{Z} = \frac{1}{\bar{Y}} - 1$$

goal: $\Pr [\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z] < \delta$

 for $\epsilon \leq 1/2$

$$|\bar{Y} - \mathbb{E}[\bar{Y}]| > \frac{\epsilon/2}{z+1}$$

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry} \quad \rightarrow \quad \boxed{\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\text{let } \hat{Z} = \frac{1}{\bar{Y}} - 1$

$$\Pr \left[\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z \right]$$

$$(\text{for } \epsilon \leq 1/2) \leq \Pr \left[\left| \bar{Y} - \mathbb{E}[\bar{Y}] \right| > \frac{\epsilon/2}{z+1} \right]$$

Chebyshev: $\leq \frac{4}{\epsilon^2} (z+1)^2 \text{Var}[\bar{Y}]$

Markov's Inequality

Markov's Inequality:

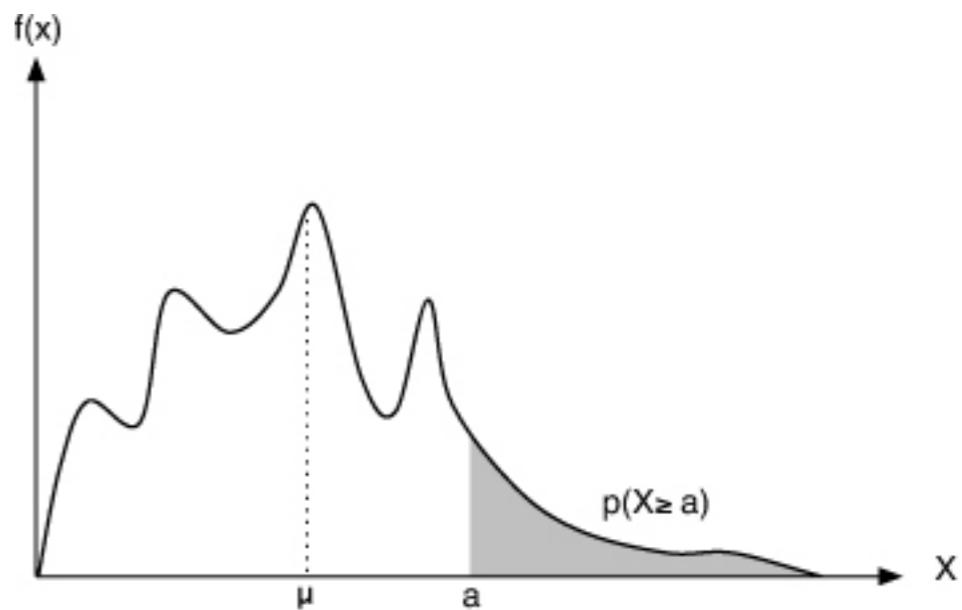
For *nonnegative* X , for any $t > 0$,

$$\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}.$$

Proof:

Let $Y = \begin{cases} 1 & \text{if } X \geq t, \\ 0 & \text{otherwise.} \end{cases}$ $\Rightarrow Y \leq \left\lfloor \frac{X}{t} \right\rfloor \leq \frac{X}{t},$

$$\Pr[X \geq t] = \mathbb{E}[Y] \leq \mathbb{E}\left[\frac{X}{t}\right] = \frac{\mathbb{E}[X]}{t}.$$



tight if we only know the expectation of X

A Generalization of Markov's Inequality

Theorem:

For any X , for $h : X \mapsto \mathbb{R}^+$, for any $t > 0$,

$$\Pr[h(X) \geq t] \leq \frac{\mathbb{E}[h(X)]}{t}.$$

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr [|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Variance:

$$\mathbf{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2$$

$$\mathbf{Var}[cX] = c^2 \mathbf{Var}[X]$$

$$\mathbf{Var} [\sum_i X_i] = \sum_i \mathbf{Var}[X_i] \quad \text{for pairwise independent } X_i$$

Chebyshev's Inequality

Chebyshev's Inequality:

For any $t > 0$,

$$\Pr [|X - \mathbf{E}[X]| \geq t] \leq \frac{\mathbf{Var}[X]}{t^2}.$$

Proof:

Apply Markov's inequality to $(X - \mathbf{E}[X])^2$

$$\Pr [(X - \mathbf{E}[X])^2 \geq t^2] \leq \frac{\mathbf{E}[(X - \mathbf{E}[X])^2]}{t^2}$$

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry} \quad \rightarrow \quad \boxed{\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}}$$

F-M estimator: $\text{let } \hat{Z} = \frac{1}{\bar{Y}} - 1$

$$\Pr \left[\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z \right]$$

$$(\text{for } \epsilon \leq 1/2) \leq \Pr \left[\left| \bar{Y} - \mathbb{E}[\bar{Y}] \right| > \frac{\epsilon/2}{z+1} \right]$$

Chebyshev: $\leq \frac{4}{\epsilon^2} (z+1)^2 \text{Var}[\bar{Y}]$

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry}$$

$$\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$$

geometry
probability $\rightarrow \Pr[Y_j \geq y] = (1-y)^z \rightarrow \text{pdf} = z(1-y)^{z-1}$

$$\mathbb{E}[Y_j^2] = \int_0^1 y^2 z(1-y)^{z-1} dy = \frac{2}{(z+1)(z+2)}$$

$$\mathbf{Var}[Y_j] = \mathbb{E}[Y_j^2] - \mathbb{E}[Y_j]^2 \leq \frac{1}{(z+1)^2}$$

$\mathbf{Var}[\bar{Y}] = \frac{1}{k^2} \sum_{j=1}^k \mathbf{Var}[Y_j] = \frac{1}{k} \mathbf{Var}[Y_j] \leq \frac{1}{k(z+1)^2}$

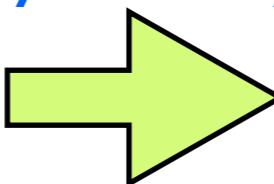
2-wise independence

$$z = |\{x_1, x_2, \dots, x_n\}|$$

For $j=1, 2, \dots, k$, hash values of h_j :

uniform independent $X_{j1}, X_{j2}, \dots, X_{jz} \in [0, 1]$

$$\left. \begin{array}{l} Y_j = \min_{1 \leq i \leq n} X_{ji} \\ \bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j \end{array} \right\} \text{symmetry}$$



$$\mathbb{E}[\bar{Y}] = \mathbb{E}[Y_j] = \frac{1}{z+1}$$

F-M estimator:

$$\text{let } \hat{Z} = \frac{1}{\bar{Y}} - 1$$

$$\Pr \left[\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z \right] \leq \frac{4}{\epsilon^2 k}$$

$$(\text{for } \epsilon \leq 1/2) \leq \Pr \left[\left| \bar{Y} - \mathbb{E}[\bar{Y}] \right| > \frac{\epsilon/2}{z+1} \right]$$

Chebyshev: $\leq \frac{4}{\epsilon^2} (z+1)^2 \text{Var}[\bar{Y}]$

$$\text{Var}[\bar{Y}] \leq \frac{1}{k(z+1)^2}$$

Input: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Output: an estimation of $z = |\{x_1, x_2, \dots, x_n\}|$

uniform independent hash functions:

$$h_1, h_2, \dots, h_k : \Omega \rightarrow [0,1] \quad Y_j = \min_{1 \leq i \leq n} h_j(x_i)$$

average-min: $\bar{Y} = \frac{1}{k} \sum_{j=1}^k Y_j$

Flajolet-Martin estimator: $\hat{Z} = \frac{1}{\bar{Y}} - 1$

UHA: Uniform Hash Assumption

$$\Pr \left[\hat{Z} > (1 + \epsilon)z \text{ or } \hat{Z} < (1 - \epsilon)z \right] \leq \frac{4}{\epsilon^2 k} \leq \delta$$

choose $k = \frac{4}{\epsilon^2 \delta}$

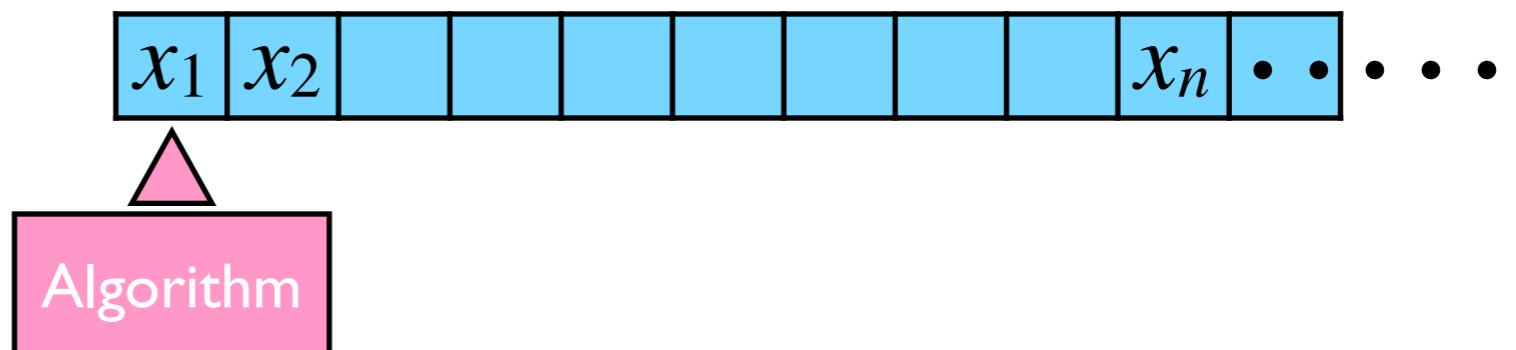
Frequency Estimation

Data: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x
within **additive error** εn .

- **data stream:** input comes one at a time



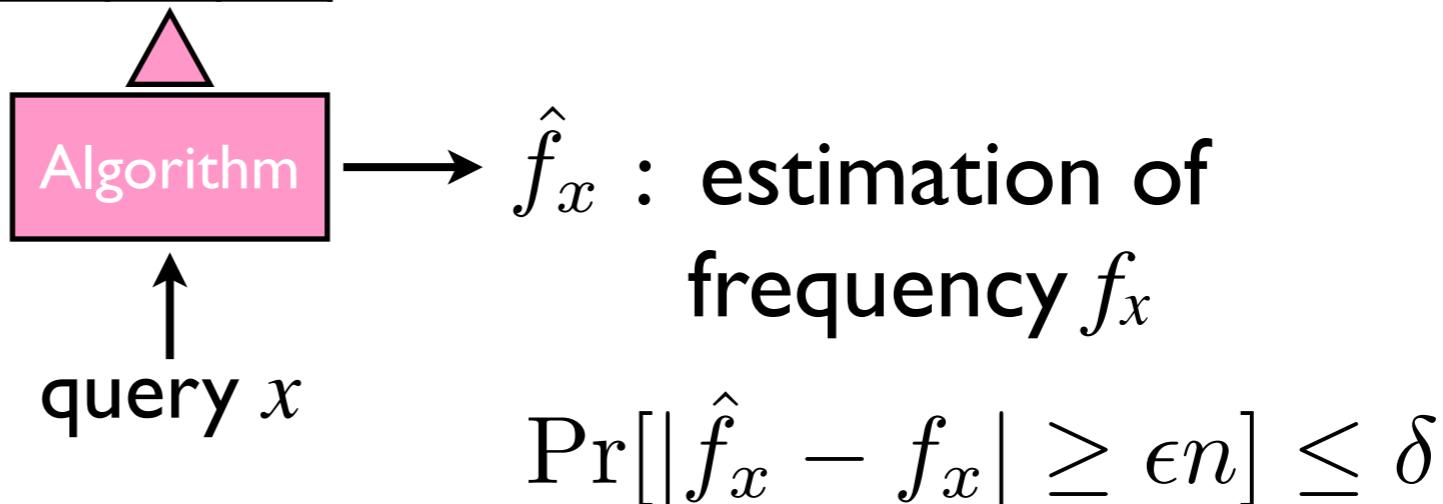
Frequency Estimation

Data: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x
within **additive error** ϵn .

- **data stream:** input comes one at a time



- **heavy hitters:** items that appears $> \epsilon n$ times

Data Structure for Set

Data: a set S of n items $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Determine whether $x \in S$.

- space cost: size of data structure (**in bits**)
 - entropy of a set: $O(n \log|\Omega|)$ bits
- time cost: time to answer a query
- **balanced tree:** $O(n \log|\Omega|)$ space, $O(\log n)$ time
- **perfect hashing:** $O(n \log|\Omega|)$ space, $O(1)$ time
- using $<$ entropy space ? a **sketch** of the set
(approximate representation)

Approximate a Set

Data: a set S of n items $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Determine whether $x \in S$.

uniform hash function $h: \Omega \rightarrow [m]$

data structure: an m -bit vector $v \in \{0, 1\}^m$

initially v is all-0;

set $v[h(x_i)] = 1$ for each $x_i \in S$;

query x : answer “yes” if $v[h(x)] = 1$;

$x \in S$: always correct

$x \notin S$: **false positive** $\Pr[v[h(x)] = 1] = 1 - (1 - 1/m)^n = 1 - e^{-n/m}$

Bloom Filters

(Bloom 1970)

Data: a set S of n items $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Determine whether $x \in S$.

uniform independent hash functions

$$h_1, h_2, \dots, h_k: \Omega \rightarrow [m]$$

data structure: an m -bit vector $v \in \{0, 1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)] = 1$ for all $j=1, \dots, k$;

query x : “yes” if $v[h_j(x)] = 1$ for all $j=1, \dots, k$;

Bloom Filters

uniform independent hash functions

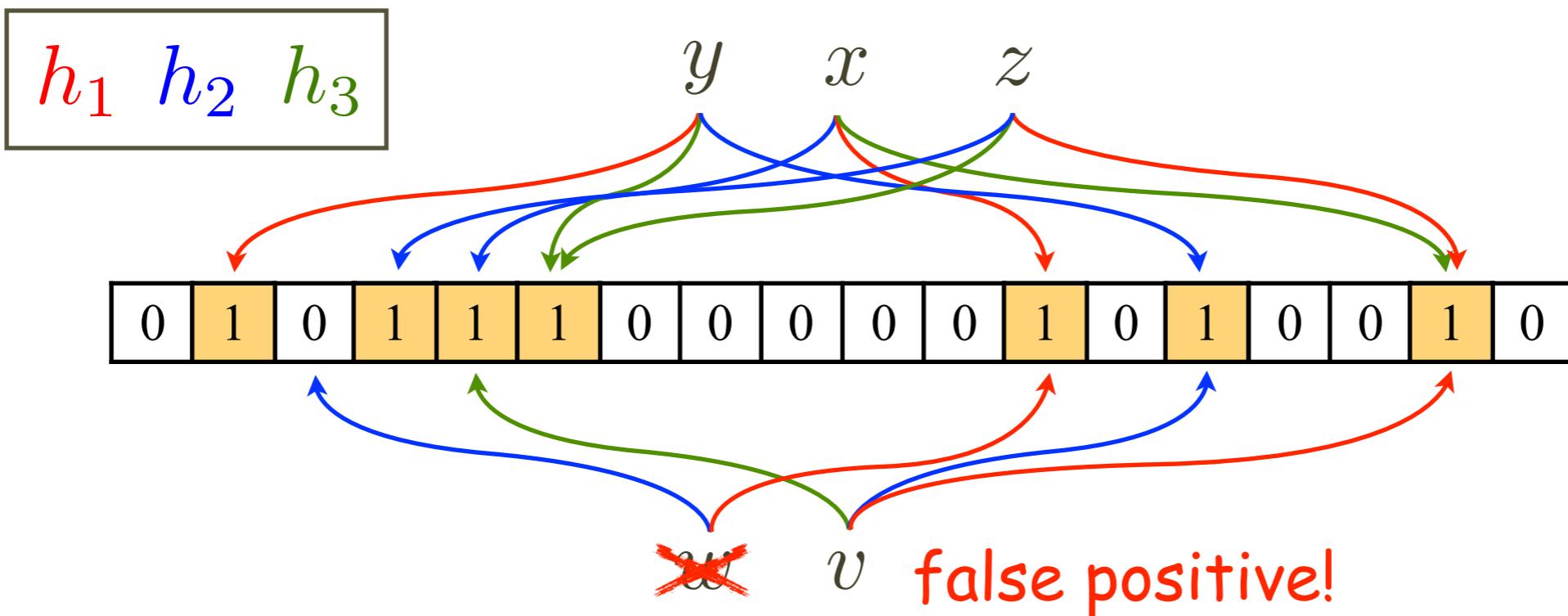
$$h_1, h_2, \dots, h_k: \Omega \rightarrow [m]$$

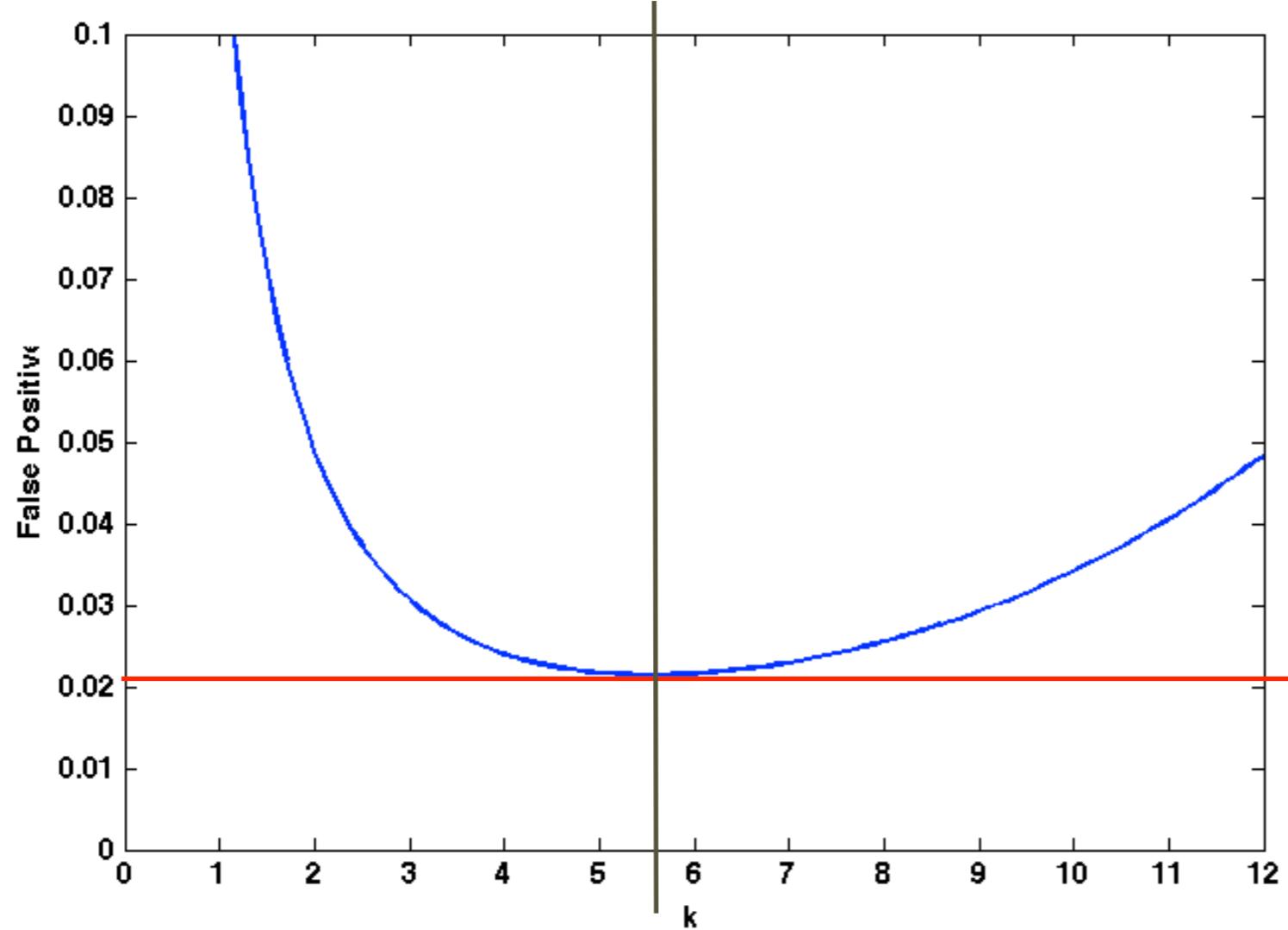
data structure: an m -bit vector $v \in \{0, 1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)] = 1$ for all $j=1, \dots, k$;

query x : “yes” if $v[h_j(x)] = 1$ for all $j=1, \dots, k$;





try: $x \in \Omega$

$0, 1}^m$

| $j=1, \dots, k;$

, ..., $k;$

$x \notin S$: **false positive**

$$\Pr[\forall 1 \leq j \leq k : v[h_j(x)] = 1]$$

$$= (\Pr[v[h_j(x)] = 1])^k = (1 - \Pr[v[h_j(x)] = 0])^k$$

$$\leq (1 - (1 - 1/m)^{kn})^k = (1 - e^{-kn/m})^k \approx (0.6185)^c$$

choose $k = \frac{m \ln 2}{n}$
 $m = cn$

Bloom Filters

data: set $S \subseteq \Omega$ of size $|S|=n$ **query:** $x \in \Omega$

uniform independent hash functions

$$h_1, h_2, \dots, h_k: \Omega \rightarrow [m]$$

data structure: an m -bit vector $v \in \{0, 1\}^m$

initially v is all-0;

for each $x_i \in S$: set $v[h_j(x_i)] = 1$ for all $j=1, \dots, k$;

query x : “yes” if $v[h_j(x)] = 1$ for all $j=1, \dots, k$;

choose $m = cn$ $k = \frac{m \ln 2}{n} = c \ln 2$

- space cost: cn bits; time cost: $c \ln 2$
- **false positive**: $< (0.6185)^c$

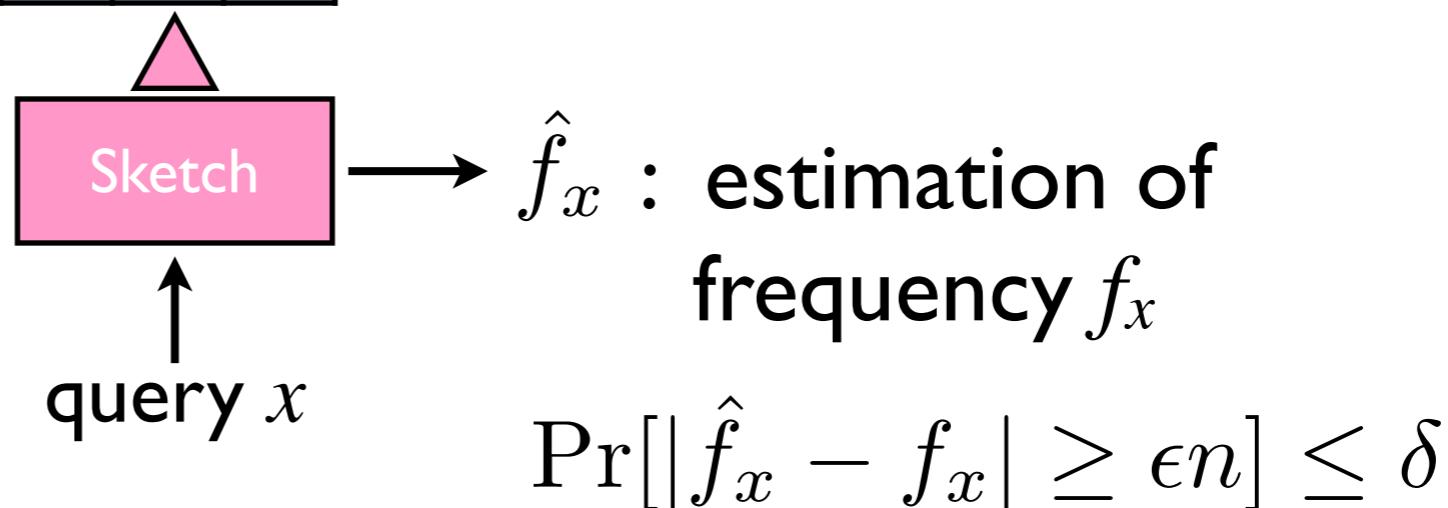
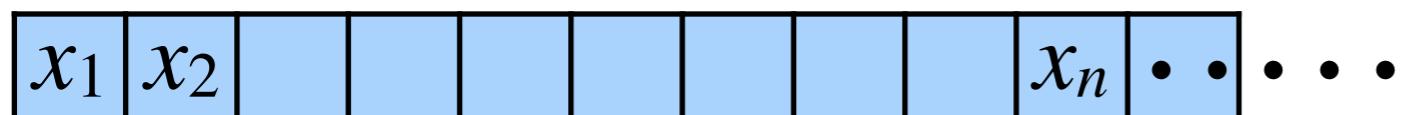
Heavy Hitters

Data: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Estimate the *frequency* $f_x = |\{i : x_i = x\}|$ of item x within **additive error** ϵn .

- **data stream:** input comes one at a time



- **heavy hitters:** items that appears $> \epsilon n$ times

Count-Min Sketch

Data: a sequence $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

Estimate the **frequency** $f_x = |\{i : x_i = x\}|$ of item x within **additive error** εn .

uniform independent hash functions

$h_1, h_2, \dots, h_k : \Omega \rightarrow [m]$

count-min sketch: CMS[k][m]

initially CMS[][] is all-0;

for each x_i and each h_j : CMS[j][$h_j(x_i)$] ++;

query x : return $\hat{f}_x = \min_{1 \leq j \leq k} \text{CMS}[j][h_j(x)]$

obviously CMS[j][$h_j(x)$] $\geq f_x$ for all $j=1,2,\dots,k$

data: $x_1, x_2, \dots, x_n \in \Omega$ **query:** $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$h_1, h_2, \dots, h_k : \Omega \rightarrow [m]$

count-min sketch: CMS[k][m]

initially CMS[][] is all-0;

for each x_i and each h_j : CMS[j][$h_j(x_i)$] ++;

query x : return $\hat{f}_x = \min_{1 \leq j \leq k} \text{CMS}[j][h_j(x)]$

for any $x \in \Omega$, for any j :

$$\text{CMS}[j][h_j(x)] = f_x + \sum_{\substack{y \in \{x_1, \dots, x_n\} \setminus \{x\} \\ h_j(y) = h_j(x)}} f_y$$

$$\mathbb{E} [\text{CMS}[j][h_j(x)]] = f_x + \sum_{y \in \{x_1, \dots, x_n\} \setminus \{x\}} f_y \Pr[h_j(y) = h_j(x)]$$

data: $x_1, x_2, \dots, x_n \in \Omega$ **query:** $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$h_1, h_2, \dots, h_k : \Omega \rightarrow [m]$

count-min sketch: CMS[k][m]

initially CMS[][] is all-0;

for each x_i and each h_j : CMS[j][$h_j(x_i)$] ++;

query x : return $\hat{f}_x = \min_{1 \leq j \leq k} \text{CMS}[j][h_j(x)]$

for any $x \in \Omega$, for any j :

$$\mathbb{E} [\text{CMS}[j][h_j(x)]] = f_x + \sum_{y \in \{x_1, \dots, x_n\} \setminus \{x\}} f_y \Pr[h_j(y) = h_j(x)]$$

$$= f_x + \frac{1}{m} \sum_{y \in \{x_1, \dots, x_n\} \setminus \{x\}} f_y \leq f_x + \frac{1}{m} \sum_{y \in \{x_1, \dots, x_n\}} f_y = f_x + \frac{n}{m}$$

biased estimator

data: $x_1, x_2, \dots, x_n \in \Omega$ **query:** $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$h_1, h_2, \dots, h_k : \Omega \rightarrow [m]$

count-min sketch: CMS[k][m]

initially CMS[][] is all-0;

for each x_i and each h_j : CMS[j][$h_j(x_i)$] ++;

query x : return $\hat{f}_x = \min_{1 \leq j \leq k} \text{CMS}[j][h_j(x)]$

$$\forall x, \forall j : \text{CMS}[j][h_j(x)] \geq f_x$$

$$\mathbb{E} [\text{CMS}[j][h_j(x)]] \leq f_x + \frac{n}{m}$$

Markov's inequality: $\Pr[\text{CMS}[j][h_j(x)] - f_x \geq \epsilon n] \leq 1/(\epsilon m)$

$$\Pr [|\hat{f}_x - f_x| \geq \epsilon n] = \Pr [\forall j : \text{CMS}[j][h_j(x)] - f_x \geq \epsilon n] \leq 1/(\epsilon m)^k$$

data: $x_1, x_2, \dots, x_n \in \Omega$ **query:** $x \in \Omega$

frequency $f_x = |\{i : x_i = x\}|$ of item x

uniform independent hash functions

$h_1, h_2, \dots, h_k : \Omega \rightarrow [m]$

count-min sketch: CMS[k][m]

initially CMS[][] is all-0;

for each x_i and each h_j : CMS[j][$h_j(x_i)$] ++;

query x : return $\hat{f}_x = \min_{1 \leq j \leq k} \text{CMS}[j][h_j(x)]$

$$\Pr \left[\left| \hat{f}_x - f_x \right| \geq \epsilon n \right] \leq 1/(\epsilon m)^k \leq \delta$$

choose $m = \lceil \frac{\epsilon}{\epsilon} \rceil$ $k = \lceil \ln \frac{1}{\delta} \rceil$

- **space cost:** $km = O\left(\frac{1}{\epsilon} \ln \frac{1}{\delta}\right)$
- **time cost** for each query: $k = O\left(\ln \frac{1}{\delta}\right)$

Set Membership

Data: a set S of n items $x_1, x_2, \dots, x_n \in \Omega$

Query: an item $x \in \Omega$

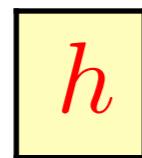
Determine whether $x \in S$.

- space cost: size of data structure (**in bits**)
 - entropy of a set: $O(n \log |\Omega|)$ bits
- time cost: time to answer a query
- **balanced tree:** $O(n \log |\Omega|)$ space, $O(\log n)$ time
- **perfect hashing:** $O(n \log |\Omega|)$ space, $O(1)$ time

Perfect Hashing

$$S = \{a, b, c, d, e, f\} \subseteq [N]$$

uniform
random



$$[N] \rightarrow [m]$$

no collision

$$\Pr[\text{perfect}] > 1/2$$

Table T :

e	b		d		f		c	a	
---	---	--	---	--	---	--	---	---	--

$$m = O(n^2)$$

Birthday
Paradox!

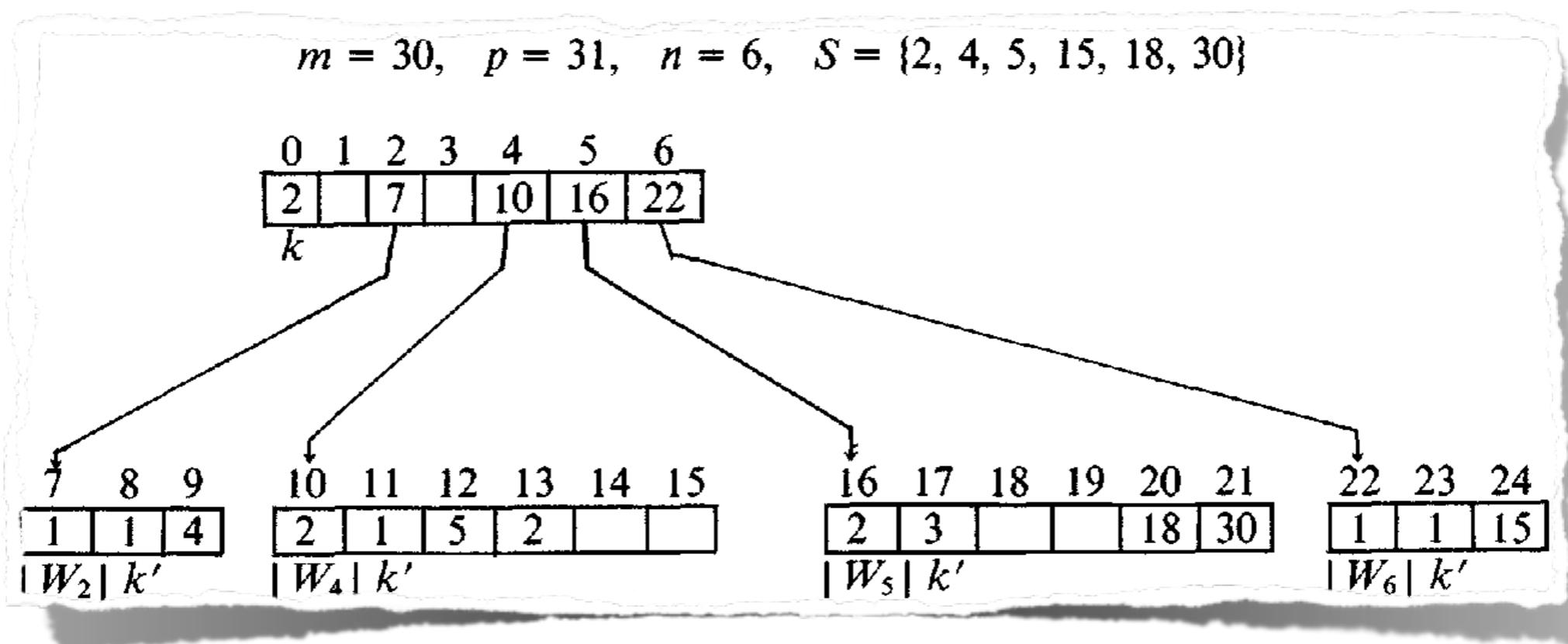
UHA: Uniform Hash Assumption

search(x): retrieve h ;

check whether $T[h(x)] = x$;

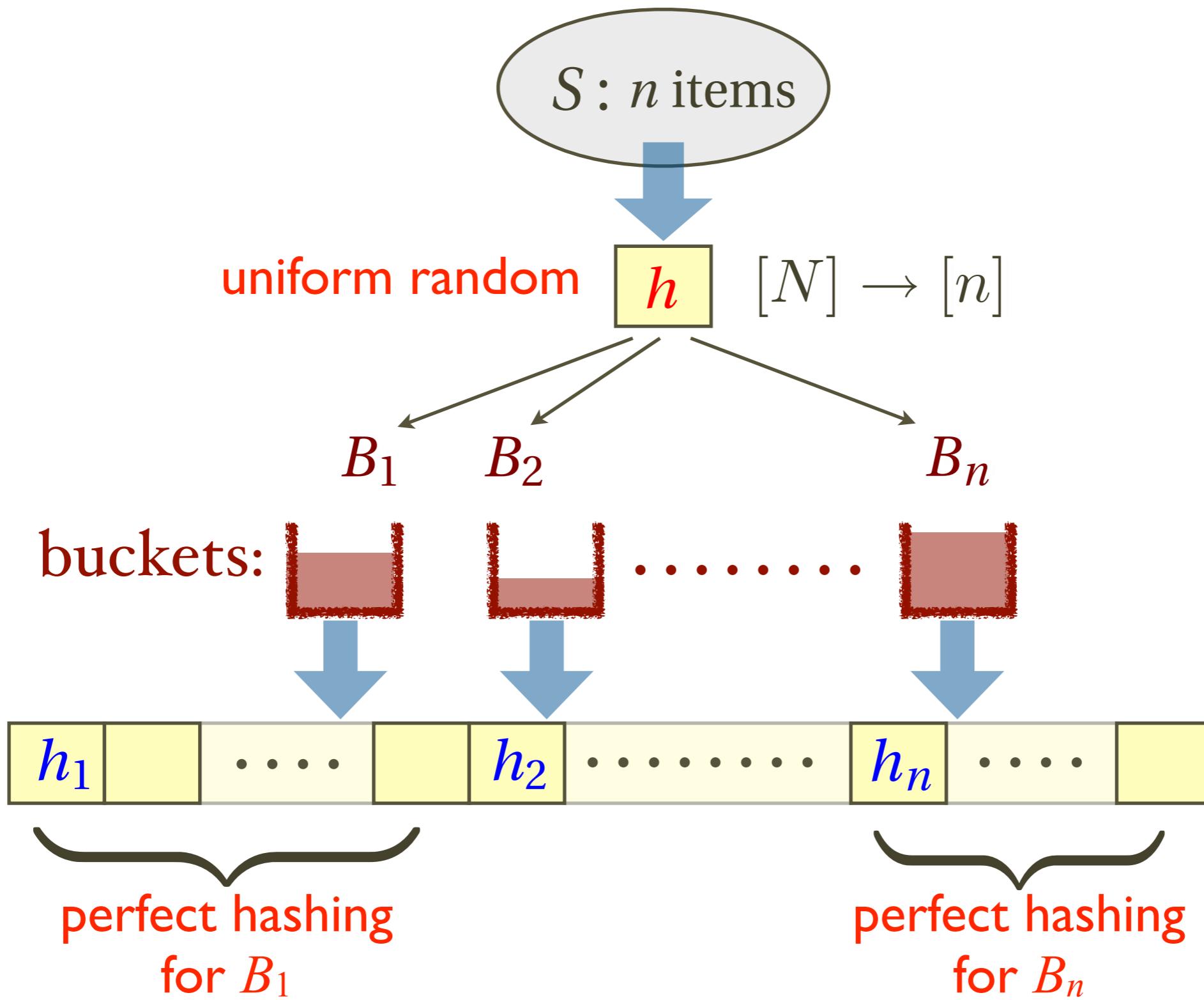
FKS Perfect Hashing

(Fredman, Komlós, Szemerédi, 1984)



- **space cost:** $O(n)$ words ($O(n \log |\Omega|)$ bits);
- **time cost:** $O(1)$ for every query in the worst case.

FKS Perfect Hashing



FKS Perfect Hashing

