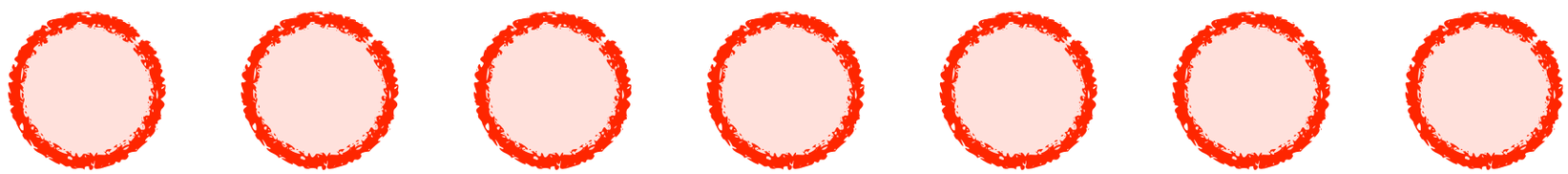
Advanced Algorithms

Balls into Bins

尹一通 Nanjing University, 2021 Fall

Balls into Bins

n balls



uniform & independent

m bins

random function $f : [n] \to [m]$

birthday, coupon collector, occupancy, ...

Random Function

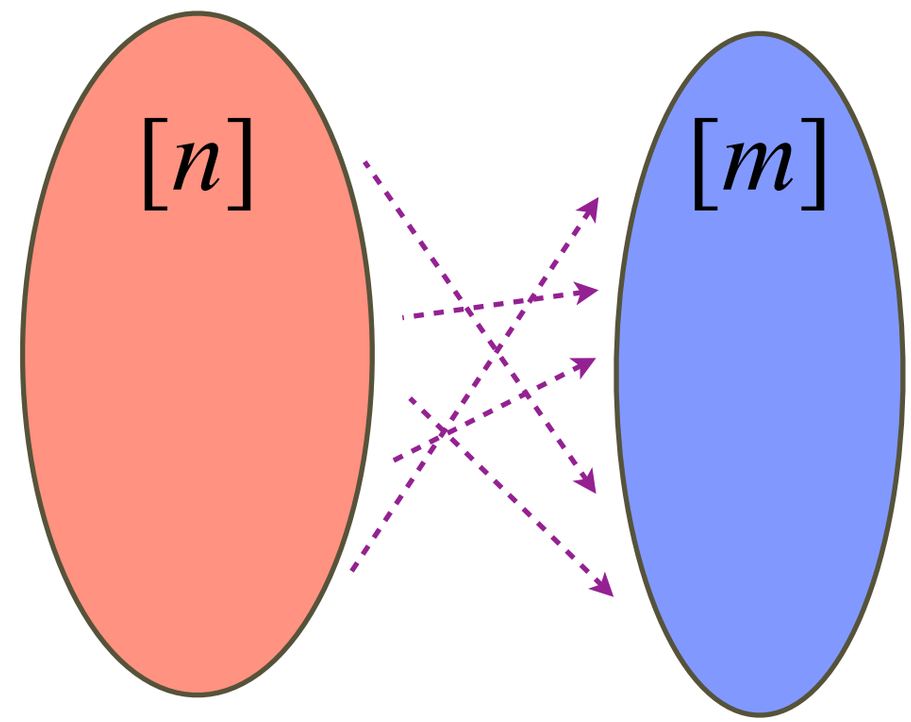
- n balls into m bins:

$$\Pr[\text{assignment}] = \frac{1}{m} \cdots \frac{1}{m} = \frac{1}{m^n}$$

- uniform random function:

$$\Pr[f] = \frac{1}{\left|[n] \to [m]\right|} = \frac{1}{m^n}$$

1-1	birthday
on-to	coupon collector
pre-image size	occupancy



uniform random function

$$f : [n] \to [m]$$

Birthday Paradox

Paradox:

(i) a statement that leads to a contradiction;

(ii) a situation which defies intuition.

In a class of m>57 students, with >99% probability, there are two students with the same birthday.

Assumption: birthdays are uniformly & independently distributed.

n balls are thrown into m bins:

event $\mathscr{E}$: each bin receives ≤ 1 balls

Birthday Paradox

n balls are thrown into m bins:
event $\mathscr{E}$: each bin receives ≤ 1 balls

$$\Pr[\mathscr{E}] = \frac{\left|[n] \xrightarrow{1-1} [m]\right|}{\left|[n] \to [m]\right|} = \frac{m(m-1)\cdots(m-n+1)}{m^n}$$

$$= \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right)$$

Birthday Paradox

n balls are thrown into m bins:
event $\mathscr{E}$: each bin receives ≤ 1 balls

Suppose that balls are thrown one-by-one:

$\Pr[\mathscr{E}] = \Pr[$all n balls are thrown into ditinct bins$]$

chain rule $= \prod_{i=1}^{n} \Pr[$the ith ball is thrown into an empty bin $|$ first $i-1$ balls are thrown into ditinct bins$]$

$$= \prod_{i=1}^{n} \left(1 - \frac{i-1}{m}\right) = \prod_{i=0}^{n-1} \left(1 - \frac{i}{m}\right)$$

Birthday Paradox

n balls are thrown into m bins:
event $\mathscr{E}$: each bin receives ≤ 1 balls

(Taylor: $1 - x \approx \mathrm{e}^{-x}$ for $x = o(1)$)

$$\Pr[\mathscr{E}] = \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right) \approx \prod_{i=0}^{n-1} \mathrm{e}^{-\frac{i}{m}} \approx \mathrm{e}^{-n^2/2m}$$

Formally: $\mathrm{e}^{-(1+o(1))n^2/2m} \leq \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right) \leq \mathrm{e}^{-(1-o(1))n^2/2m}$

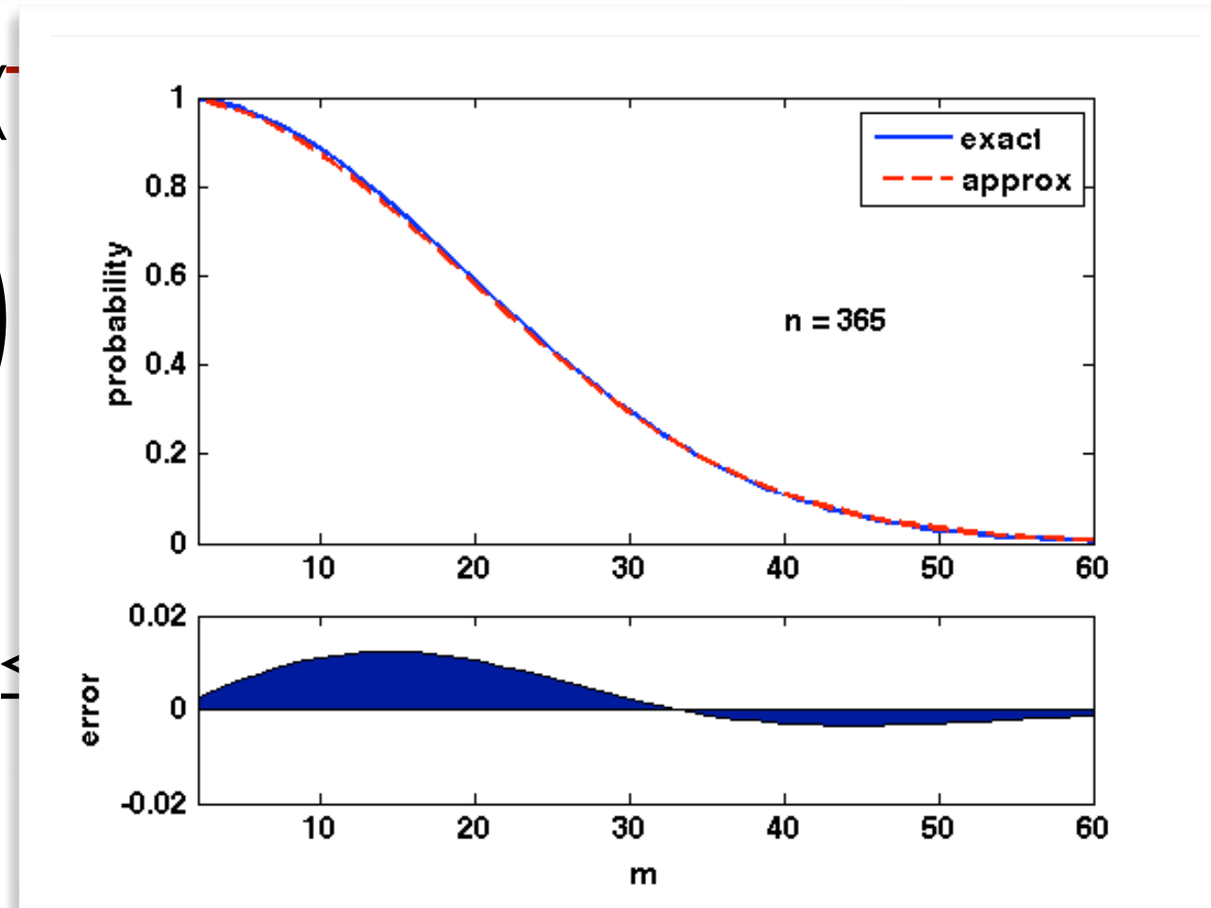
(assuming $n \ll m$)

when $n = \sqrt{2m \ln \frac{1}{p}} \implies \Pr[\mathscr{E}] = (1 \pm o(1))p$

Birthday Paradox

n balls are thrown into m bins:
event $\mathscr{E}$: each bin receives ≤ 1 balls

$$\Pr[\mathscr{E}] = \prod_{i=0}^{n-1}\left(1 - \frac{i}{m}\right)$$



Formally: $e^{-(1+o(1))n^2/2m}$

(assuming $n \ll m$)

when $n = \sqrt{2m \ln \frac{1}{p}} \implies \Pr[\mathscr{E}] = (1 \pm o(1))p$

Data Structure for Set

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$
Query: an item $x \in U$

Determine whether $x \in S$.

- **Space cost**: size of data structure (in bits)
  - entropy of a set: $\mathrm{O}(n \log N)$ bits (when $N \gg n$)
- **Time cost**: time to answer a query (in memory accesses)
- Balanced tree: $\mathrm{O}(n \log N)$ space, $\mathrm{O}(\log n)$ time
- Perfect hashing: $\mathrm{O}(n \log N)$ space, $\mathrm{O}(1)$ time

Perfect Hashing

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

uniform random $\boxed{h}$ $[N] \to [m]$

no collision

$\Pr[\textit{perfect}] \approx e^{-n^2/2m} > 1/2$

$m = n^2$

Birthday

Table T:

e	b		d		f		c	a	

SUHA: Simple Uniform Hash Assumption

Query(x):

retrieve hash function h;

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

Universal Hashing

Universal Hash Family (Carter and Wegman 1979):

A family $\mathscr{H}$ of hash functions in $U \to [m]$ is **k-universal** if for any distinct $x_1, \ldots, x_k \in U$,

$$\Pr_{h \in \mathscr{H}} \left[h(x_1) = \cdots = h(x_k) \right] \leq \frac{1}{m^{k-1}}.$$

Moreover, $\mathscr{H}$ is **strongly k-universal** (k-wise independent) if for any distinct $x_1, \ldots, x_k \in U$ and any $y_1, \ldots, y_k \in [m]$,

$$\Pr_{h \in \mathscr{H}} \left[\bigwedge_{i=1}^{k} h(x_i) = y_i \right] = \frac{1}{m^k}.$$

k-Universal Hash Family

hash functions $h : U \to [m]$

- Linear congruential hashing:
  - Represent $U \subseteq \mathbb{Z}_p$ for sufficiently large prime p
  - $h_{a,b}(x) = ((ax + b) \bmod p) \bmod m$
  - $\mathscr{H} = \left\{ h_{a,b} \mid a \in \mathbb{Z}_p \backslash \{0\}, b \in \mathbb{Z}_p \right\}$

Theorem:
The linear congruential family $\mathscr{H}$ is 2-wise independent.

- Degree-k polynomial in finite field with random coefficients
- Hashing between binary fields: $GF(2^w) \to GF(2^l)$

$$h_{a,b}(x) = (\texttt{a*x+b})\texttt{>>}(\texttt{w-l})$$

Birthday Paradox (pairwise independence)

n balls are thrown into m bins: by 2-universal hashing
event $\mathscr{E}$: each bin receives ≤ 1 balls

- Location of n balls: $X_1, X_2, \ldots, X_n \in [m]$
- Total # of collisions:

$$Y = \sum_{i<j} I[X_i = X_j]$$

- Linearity of expectation:

$$\mathbb{E}[Y] = \sum_{i<j} \Pr[X_i = X_j] \leq \binom{n}{2} \frac{1}{m}$$

2-universal

when
$n \leq \sqrt{2m\epsilon}$

- Markov's inequality: $\Pr[\neg \mathscr{E}] = \Pr[Y \geq 1] \leq \mathbb{E}[Y] \leq \epsilon$

Perfect Hashing

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

2-universal $\boxed{h}$ $[N] \to [m]$

$$\Pr[\textit{imperfect}] = \frac{n(n-1)}{2m}$$

Table T:

e	b		d		f		c	a	

m

For 2-universal family $\mathscr{H}$ from $[N]$ to $[m]$, if $m > \binom{n}{2}$, for any $S \subseteq [N]$ of size n, there is an $h \in \mathscr{H}$ that cause no collisions over S.

Query(x):

retrieve hash function h;

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

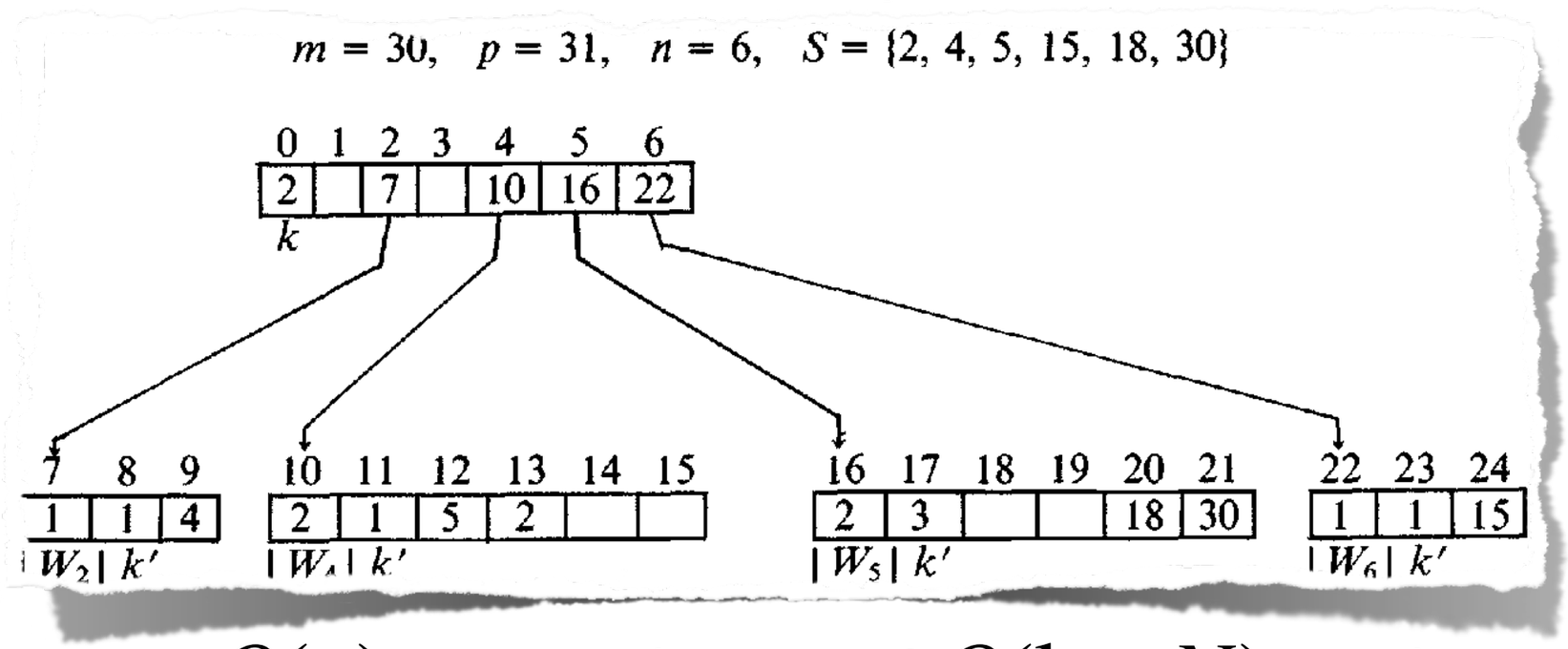
FKS Perfect Hashing

(Fredman, Komlós, Szemerédi, 1984)

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$

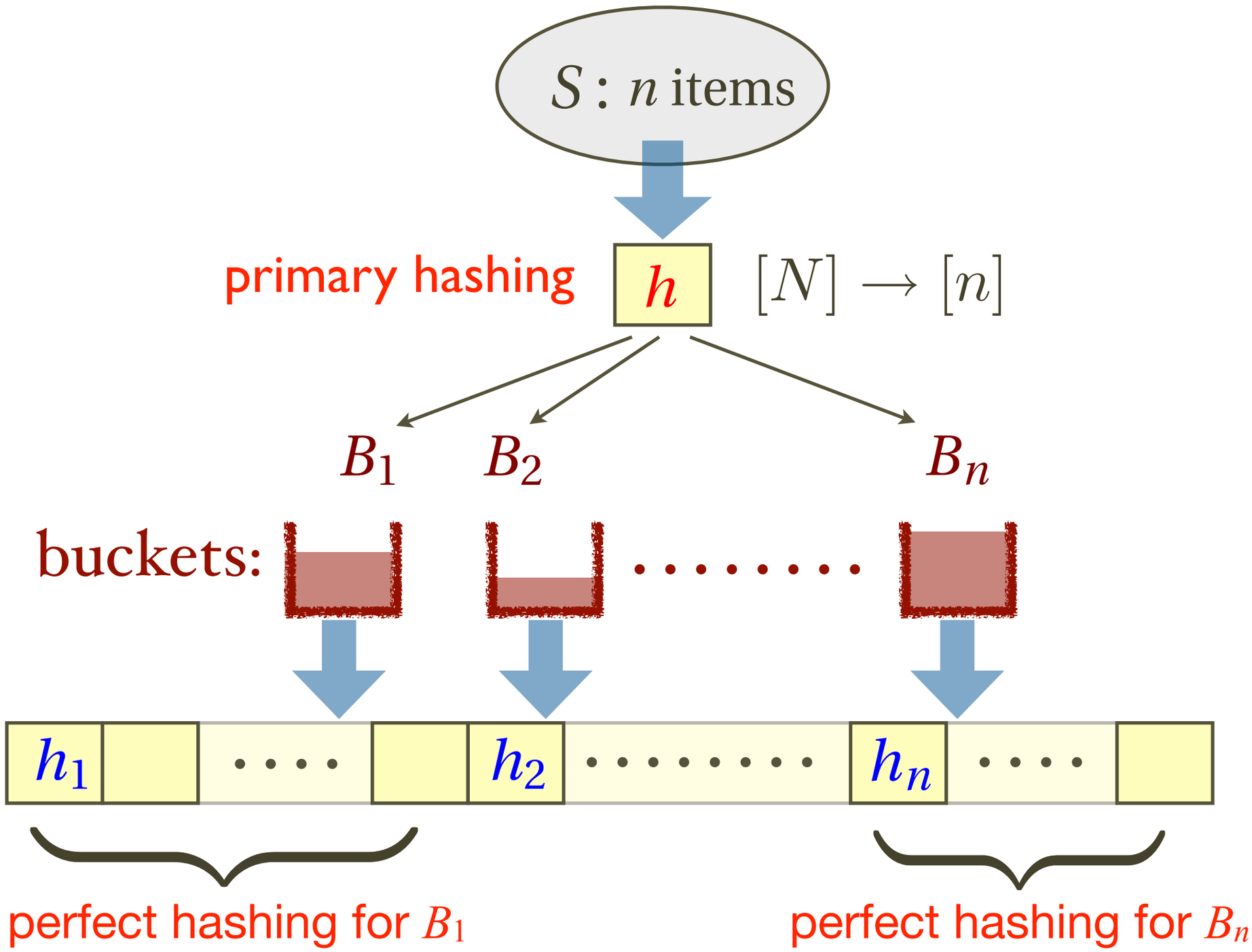
Query: an item $x \in U$

Determine whether $x \in S$.



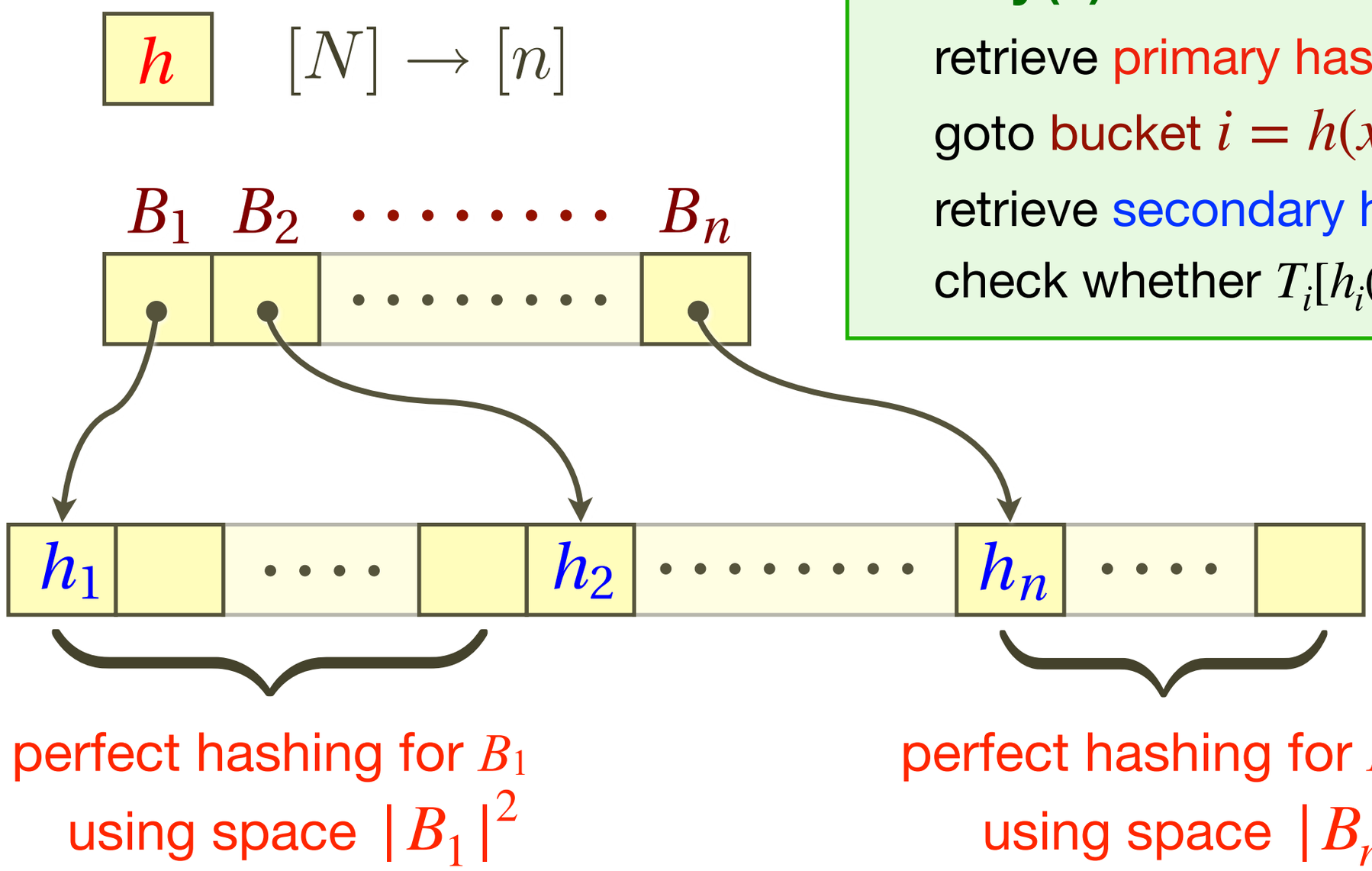
- **Space cost:** $O(n)$ words (each of $O(\log N)$ bits)
- **Time cost:** $O(1)$ for each query in the worst case

FKS Perfect Hashing



FKS Perfect Hashing

Set $S \subseteq [N]$ of size n



Query(x):

retrieve primary hash h;

goto bucket $i = h(x)$;

retrieve secondary hash h_i;

check whether $T_i[h_i(x)] = x$;

- $\exists\, h_1, \ldots, h_n$ from 2-universal family s.t. h_i is perfect for B_i for all i

Collision Number

n balls are thrown into m bins by 2-universal hashing

- Location of n bins: $X_1, X_2, \ldots, X_n \in [m]$

Collision #: $Y = \sum_{i<j} I[X_i = X_j]$

- Linearity of expectation:

$$\mathbb{E}[Y] = \sum_{i<j} \Pr[X_i = X_j] \leq \binom{n}{2} \frac{1}{m}$$

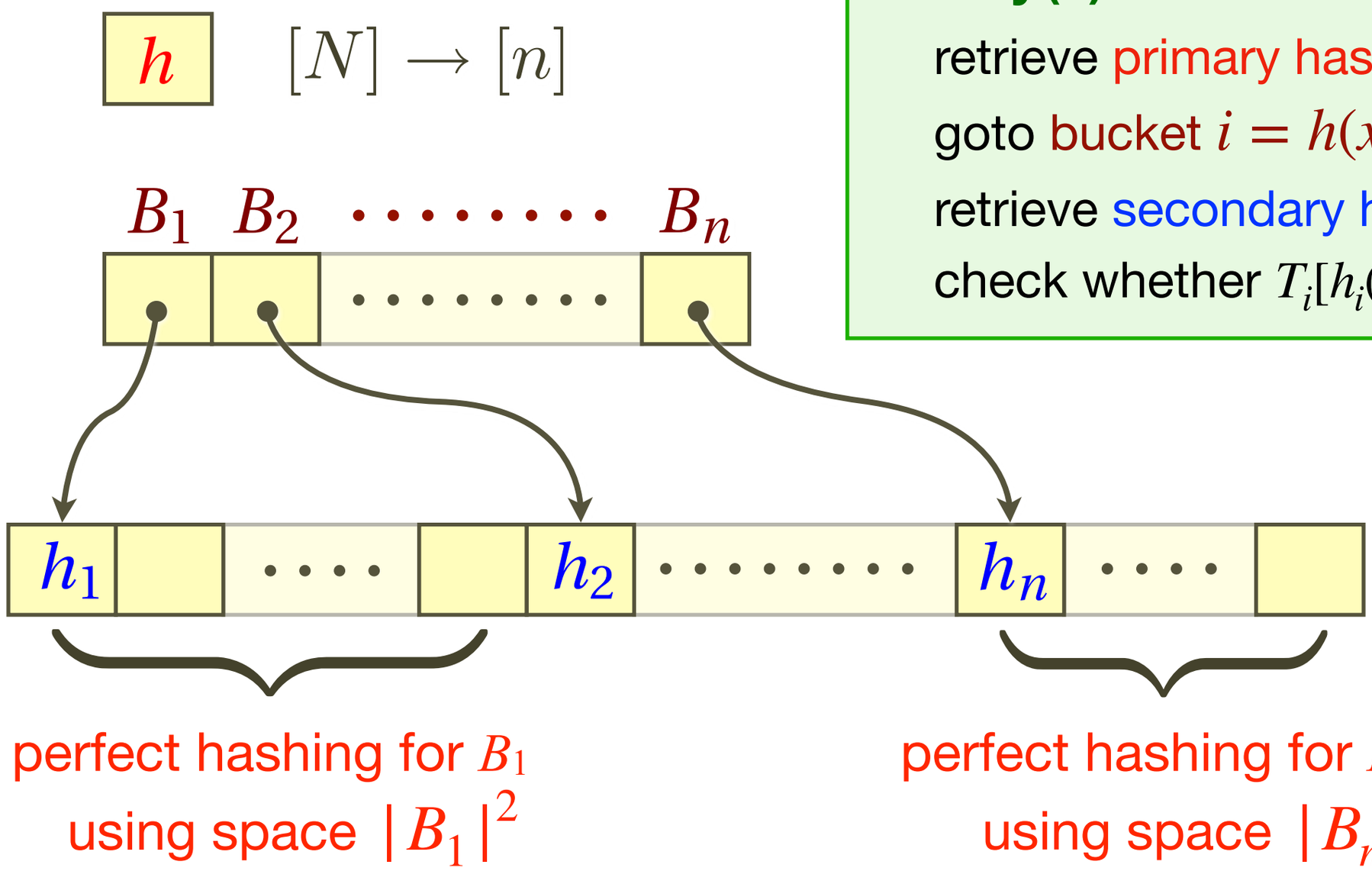
2-universal

- Size of the i-th bin: $|B_i|$

$$Y = \sum_{i=1}^{n} \binom{|B_i|}{2} = \frac{1}{2} \sum_{i=1}^{n} |B_i|(|B_i| - 1) \implies \mathbb{E}\left[\sum_{i=1}^{n} |B_i|^2\right] = \frac{n(n-1)}{m} + n$$

FKS Perfect Hashing

Set $S \subseteq [N]$ of size n



Query(x):

retrieve primary hash h;

goto bucket $i = h(x)$;

retrieve secondary hash h_i;

check whether $T_i[h_i(x)] = x$;

- $\exists h$ from a 2-universal family s.t. the total space cost is O(n)

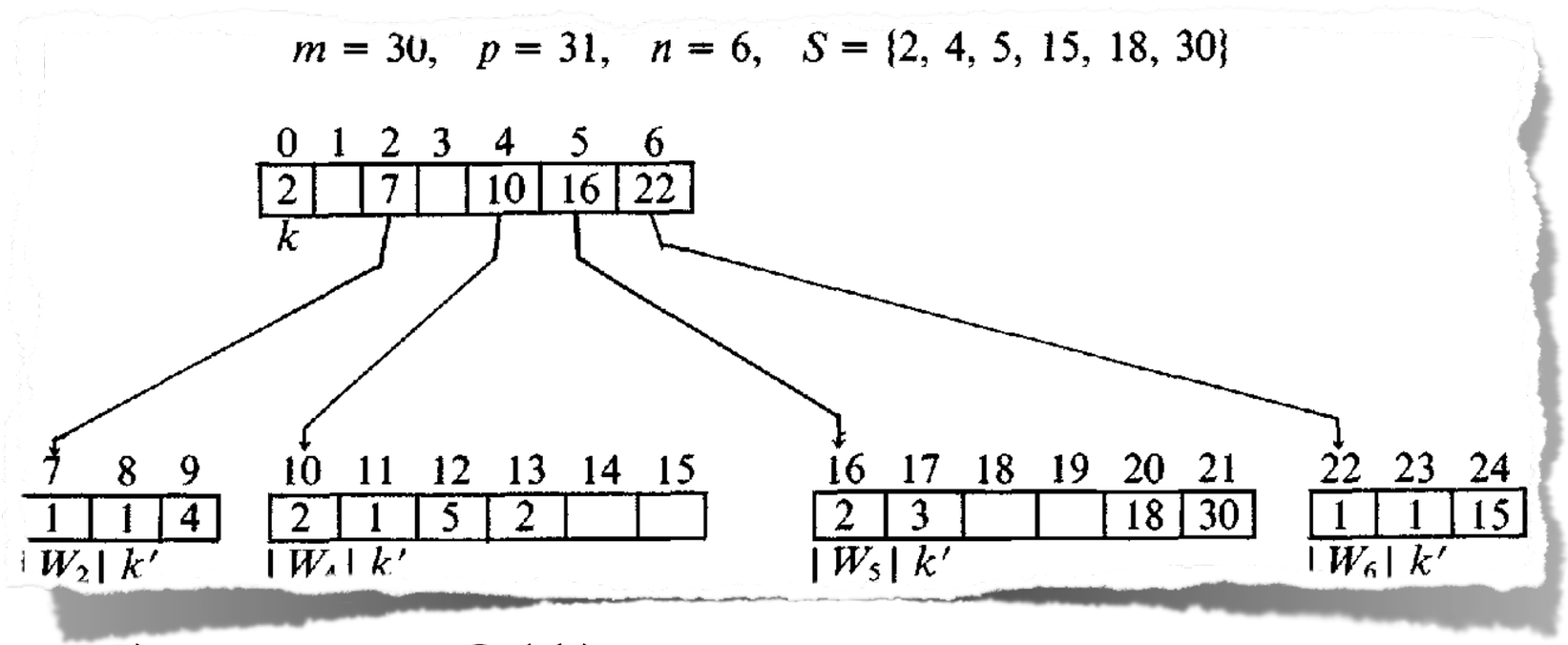
FKS Perfect Hashing

(Fredman, Komlós, Szemerédi, 1984)

Data: a set S of n items $x_1, x_2, \ldots, x_n \in U = [N]$

Query: an item $x \in U$

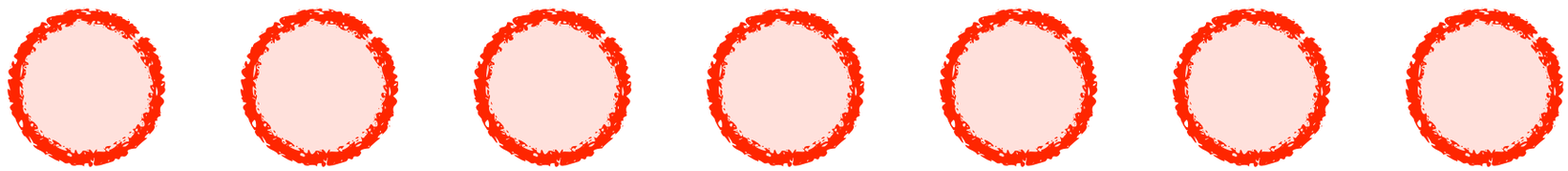
Determine whether $x \in S$.



- $O(n \log N)$ space, $O(1)$ time in the worst case
- Dynamic version: [Dietzfelbinger, Karlin, Mehlhorn, Heide, Rohnert, Tarjan, 1984]

Balls into Bins

(Coupon Collector)



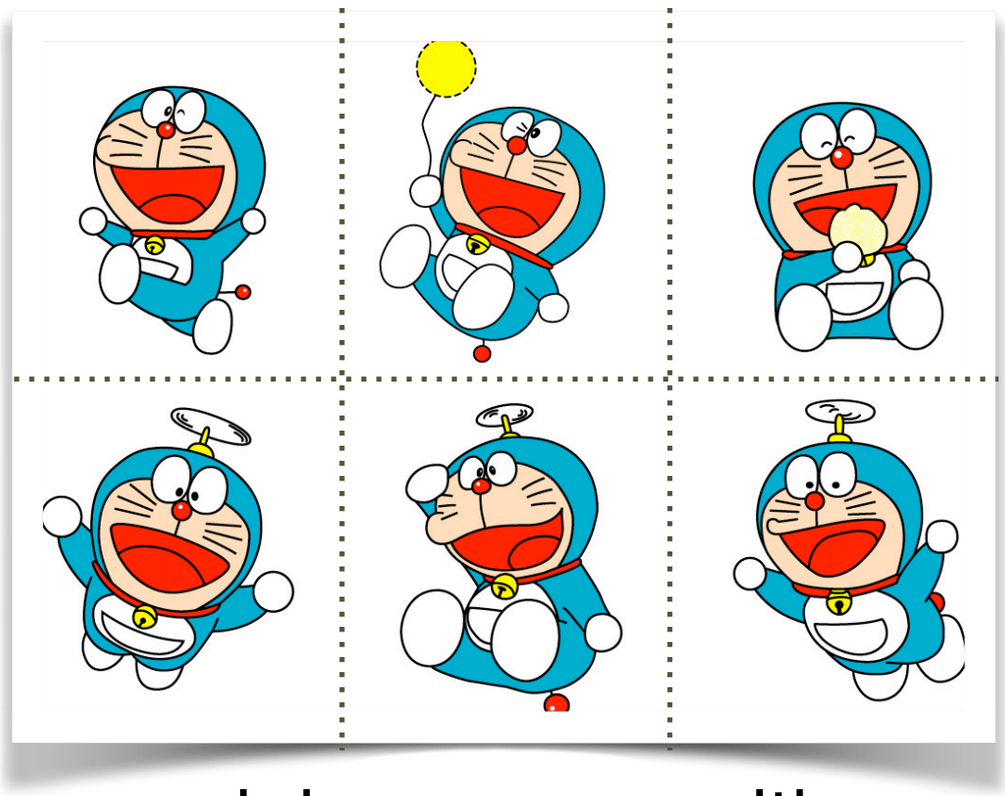
uniform & independent

n bins

surjection (cover all bins)

Coupon Collector

coupons in cookie box



each box comes with a uniformly random coupon

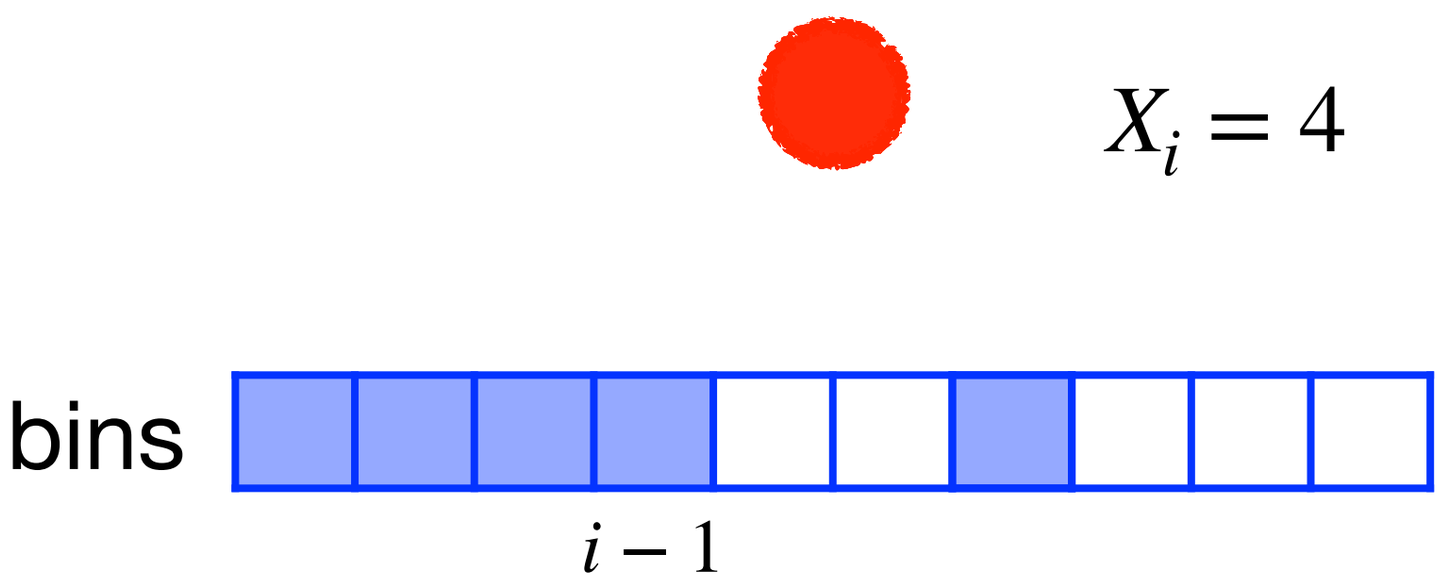
number of boxes bought to collect all n coupons

number of balls thrown to cover all n bins

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

$$X = \sum_{i=1}^{n} X_i$$



X_i is geometric!

with $p_i = 1 - \dfrac{i-1}{n}$

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n - i + 1}$$

Coupon Collector

X :	number of balls thrown to make all the n bins nonempty
X_i :	number of balls thrown while there are exactly $(i\text{-}1)$ nonempty bins

$$X = \sum_{i=1}^{n} X_i$$

$$\mathbb{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

linearity of expectations

$$= \sum_{i=1}^{n} \frac{n}{n-i+1}$$

$$= n \sum_{i=1}^{n} \frac{1}{i}$$

Harmonic number

$$= nH(n)$$

expected $n \ln n + O(n)$ balls

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

$$\Pr[\, X \geq n \ln n + cn \,] \leq \mathrm{e}^{-c}$$

Proof: For one bin, it misses all balls with probability

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} = \left(1 - \frac{1}{n}\right)^{n(\ln n + c)}$$

$$< \mathrm{e}^{-(\ln n + c)}$$

$$< \frac{1}{n\mathrm{e}^{c}}$$

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

$$\Pr[\, X \geq n \ln n + cn \,] \leq \mathrm{e}^{-c}$$

Proof: For one bin, it misses all balls with probability

$$< \frac{1}{n\mathrm{e}^{c}}$$

union bound!

$$\Pr[\, \exists \text{ a bin misses all balls} \,] \leq n \Pr[\, \text{first bin misses all bins} \,]$$

$$< \mathrm{e}^{-c}$$

Coupon Collector

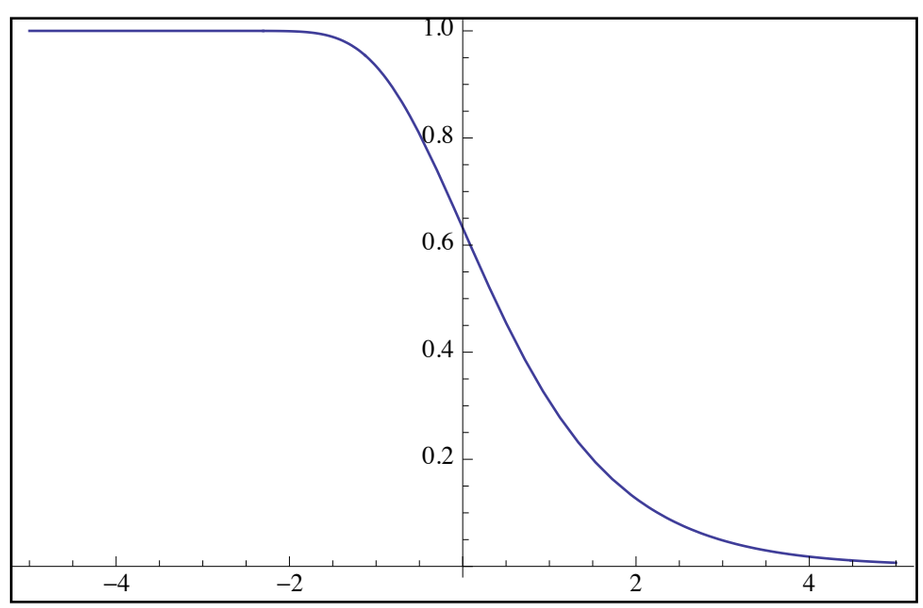
X : number of balls thrown to make all the n bins nonempty

Theorem: For $c > 0$,

$$\Pr[\, X \geq n \ln n + cn \,] \leq \mathrm{e}^{-c}$$

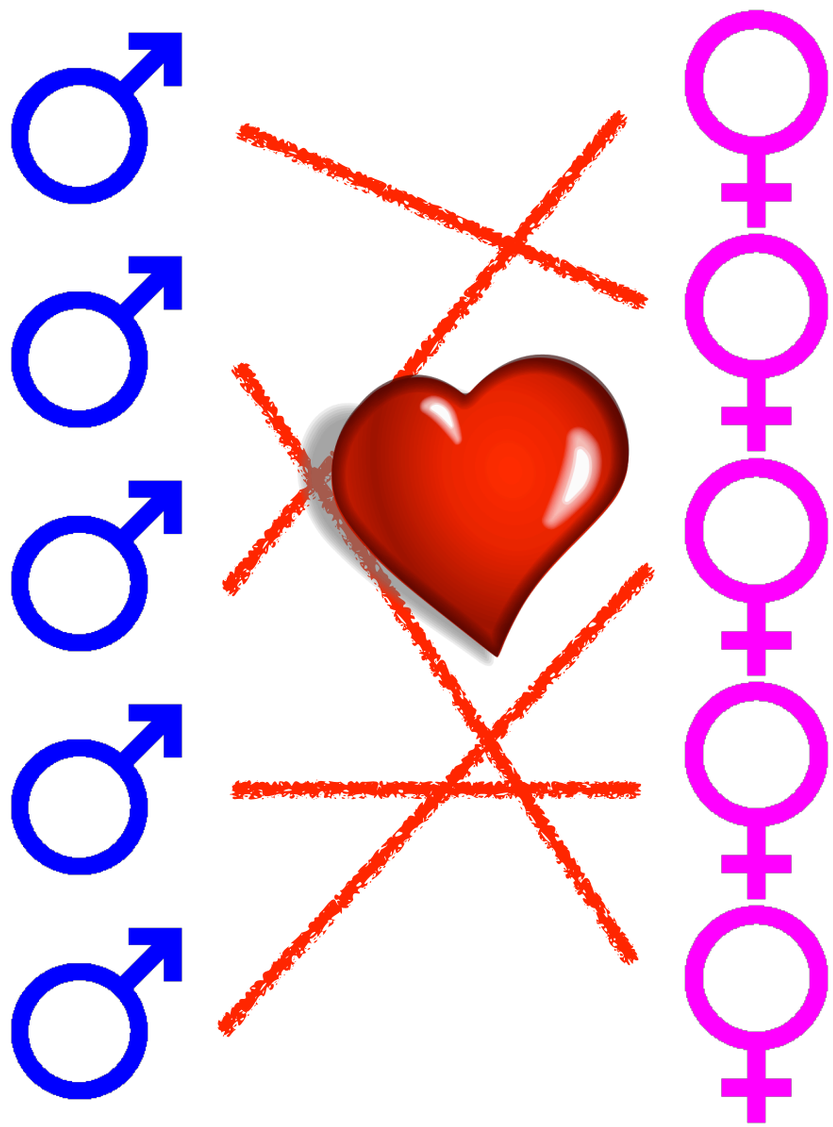
a sharp threshold:

$$\lim_{n \to \infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$



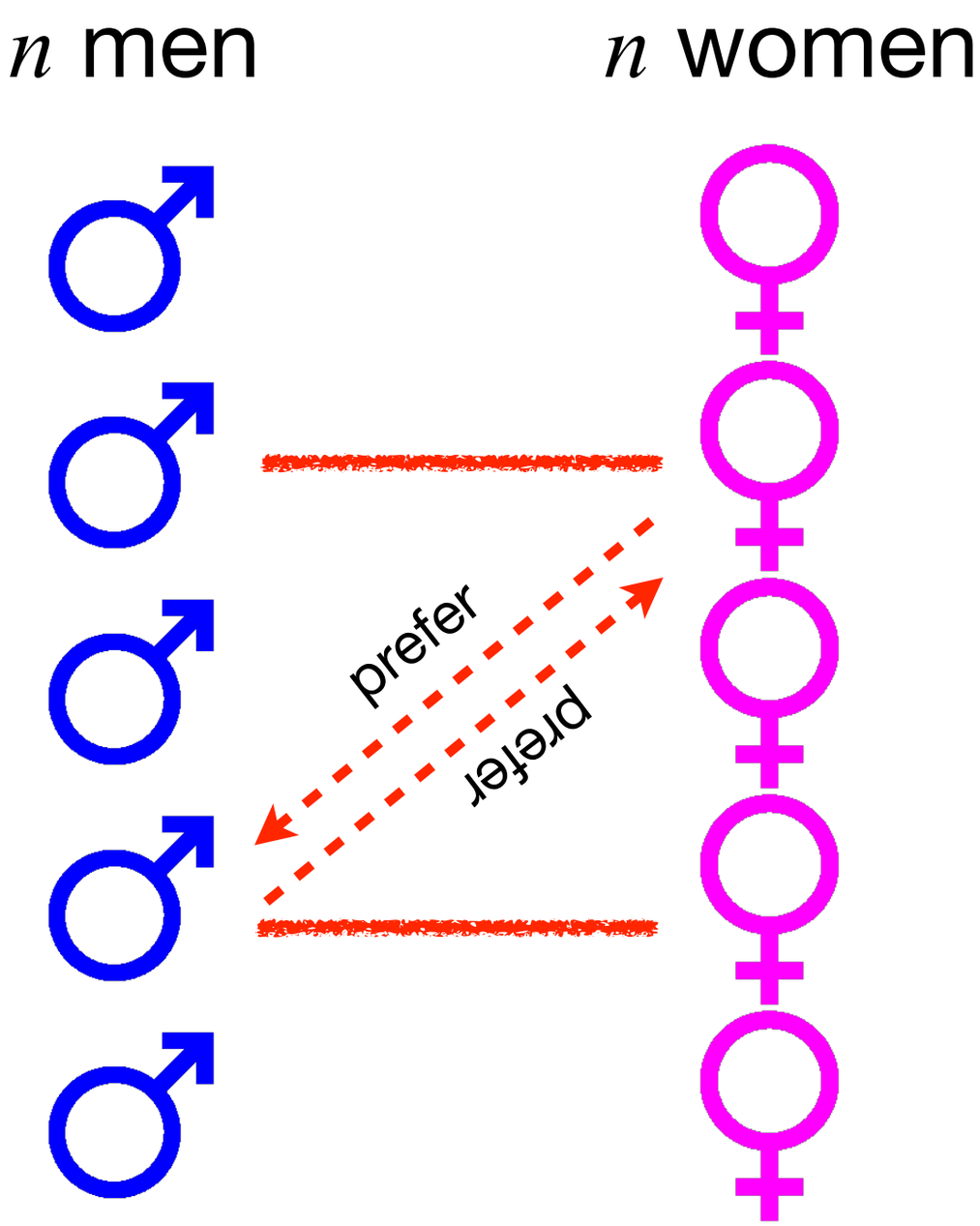
Stable Matching

n men n women



- each man has a preference order of the n women;
- each woman has a preference order of the n men;
- solution: n couples
- Marriages are stable!

Stable Matching



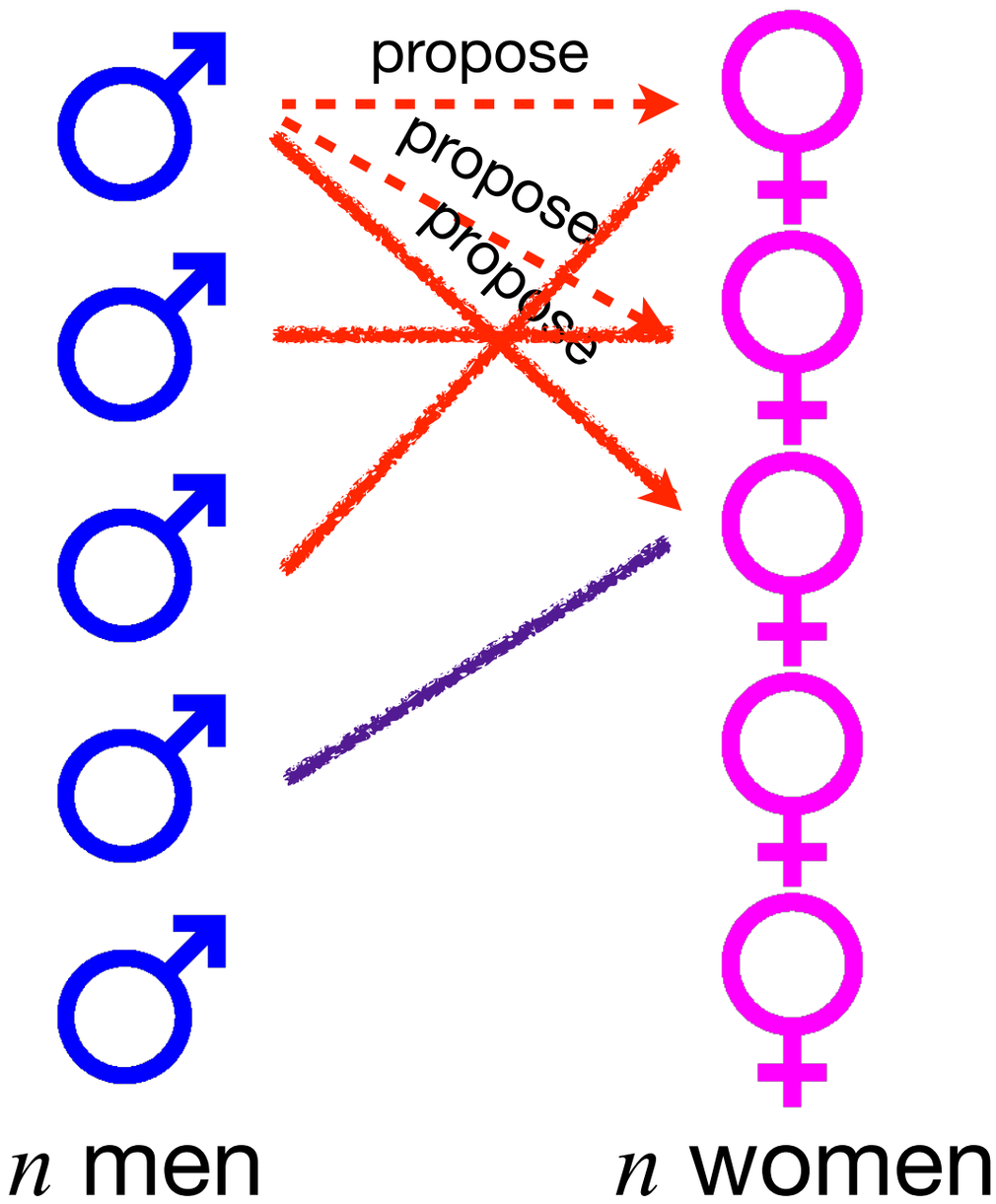
unstable (blocking pair):
a man and a woman, who prefer each other to their current partners

stable: no blocking pairs

local optimum
fixed point
equilibrium
deadlock

Proposal Algorithm

(Gale-Shapley 1962)



Single man:
propose to the most preferable women who has not rejected him

Woman:
upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Proposal Algorithm

(Gale-Shapley 1962)

- woman: once got married always married

(will only switch to better men!)

- man: will only get worse ...
- once all women are married, the algorithm terminates, and the marriages are stable
- total number of proposals:

$$\leq n^2$$

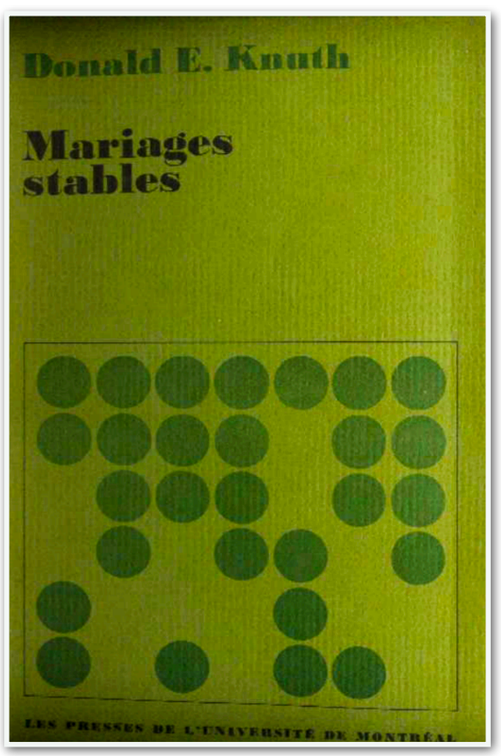
Single man:
propose to the most preferable women who has not rejected him

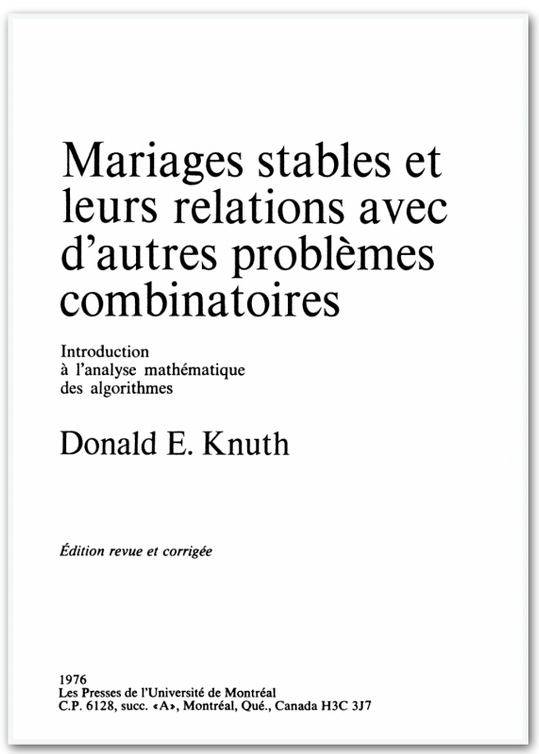
Woman:
upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Average-Case Performance

(Knuth 1976)

- Every man/woman has a uniform random permutation as preference list
- Expected total number of proposals?





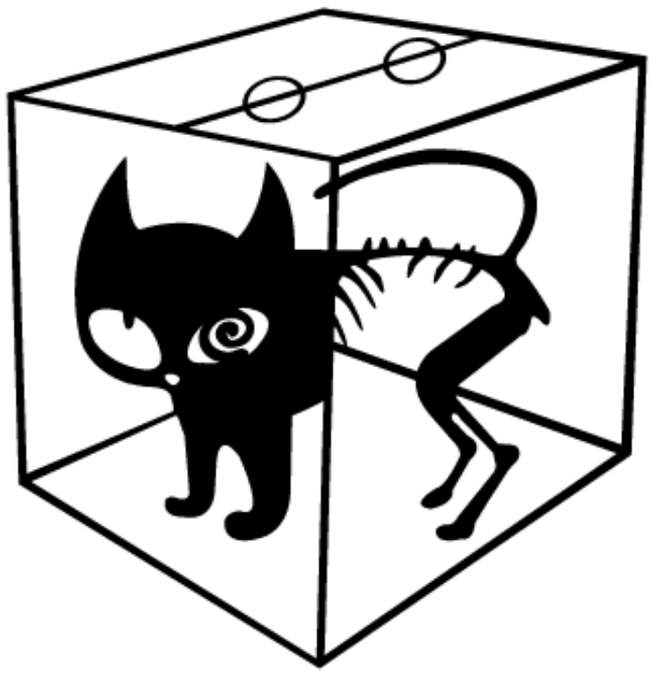
Single man:
propose to the most preferable women who has not rejected him

Woman:
upon received a proposal: accept if she's single or married to a less preferable man (divorce!)

Principle of Deferred Decisions

Principle of deferred decision

The decision of random choice in the random input is deferred to the running time of the algorithm.

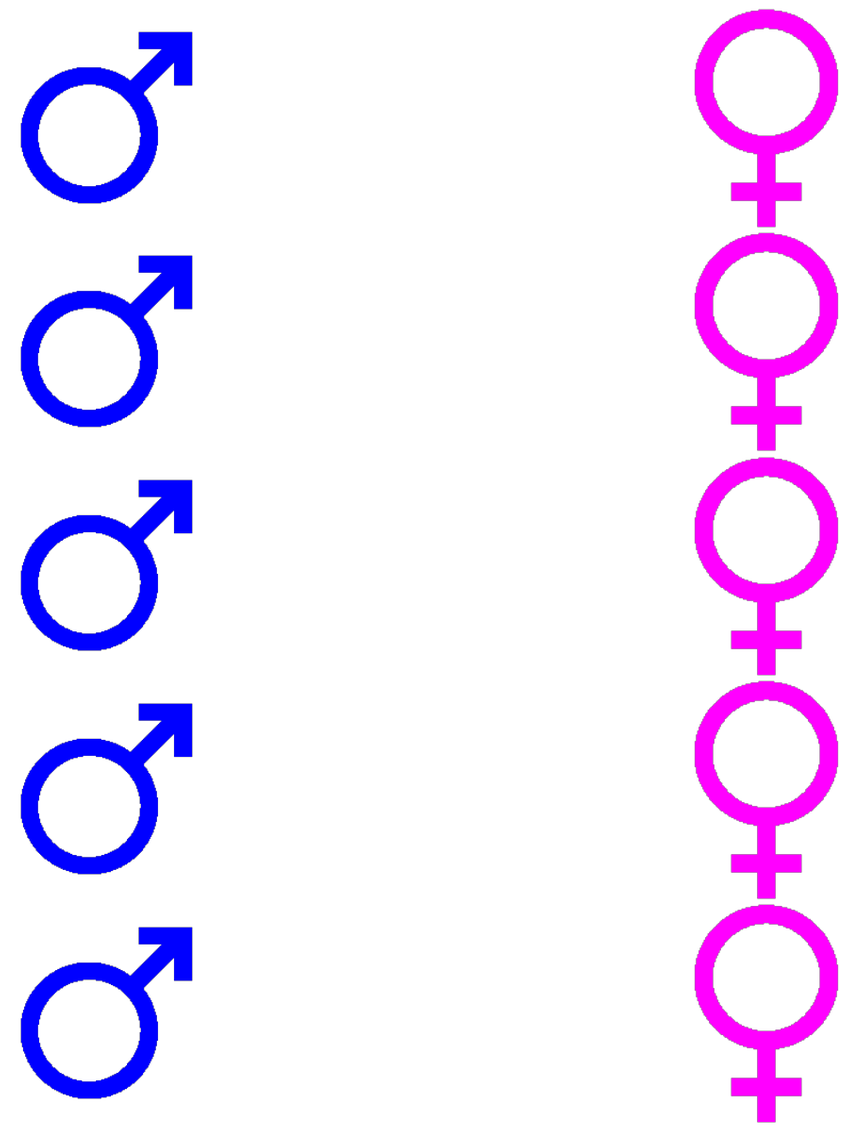


Principle of Deferred Decisions

proposing in the order of a uniformly random permutation

at each time, proposing to a uniformly random woman who has not rejected him

decisions of the inputs are deferred to the time when Alg accesses them



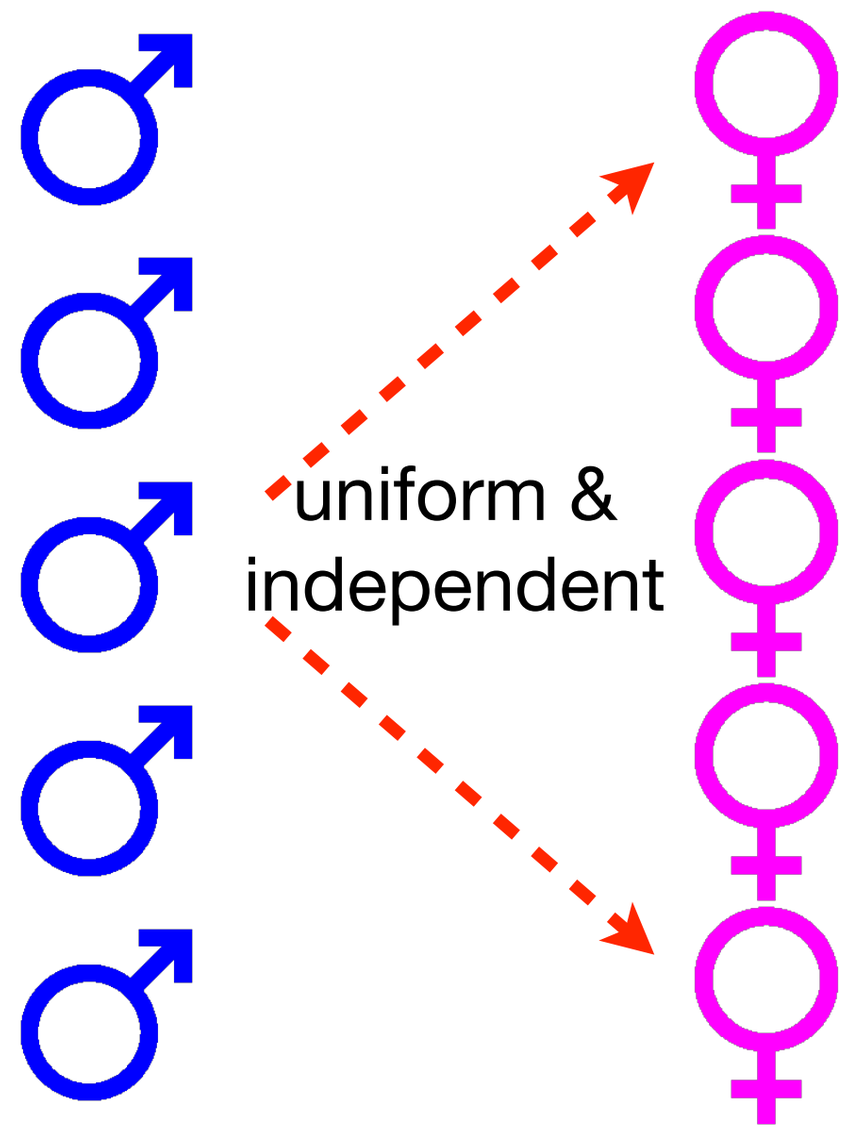
Stochastic Domination

at each time, proposing to a uniformly random woman who has not rejected him

$\wedge\!|$

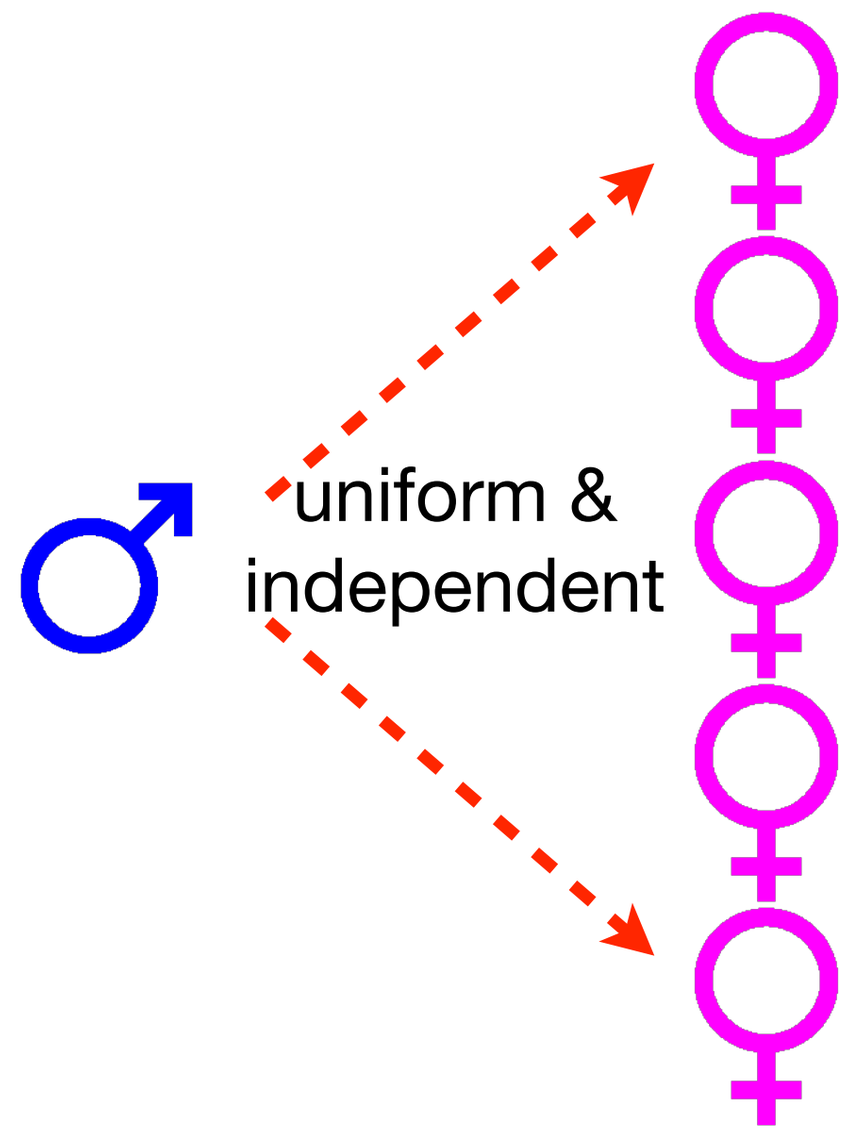
at each time, proposing to a uniformly & independently random woman

the man forgot who had rejected him (!)



Average-Case Performance

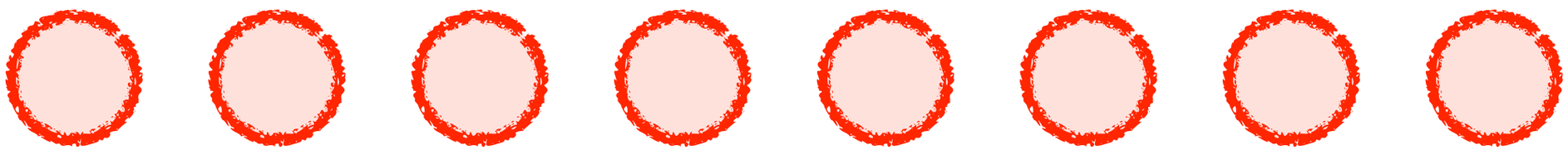
- uniformly and independently proposing to n women
- Alg stops once all women got proposed.
- Coupon collector!
- Expected $n \ln n + O(n)$ proposals.



Balls into Bins

(Occupancy Problem)

m balls



n bins

loads of bins

$X_1 \quad X_2 \quad X_3 \quad \cdots \quad \cdots \quad X_n$

maximum load?

Occupancy Problem

m balls are thrown into n bins.

X_i : number of balls in the i-th bin

$$\sum_{i=1}^{n} X_i = m \xrightarrow{\text{linearity of expectation}} \sum_{i=1}^{n} \mathbb{E}[X_i] = m$$

By symmetry: $X_1, \ldots, X_n$ are identically distributed

$$\forall i : \mathbb{E}[X_i] = \frac{m}{n}$$

$$\boxed{\max_{1 \le i \le n} \mathbb{E}[X_i] = \frac{m}{n}}$$

Occupancy Problem

m balls are thrown into n bins.

X_i : number of balls in the i-th bin

$$\max_{1 \le i \le n} \mathbb{E}[X_i] = \frac{m}{n}$$

Theorem: When $m = n$, the maximum load

$$\max_{1 \le i \le n} X_i = O\left(\frac{\log n}{\log \log n}\right) \textit{ w.h.p.}$$

w.h.p. (with high probability): with probability $1 - O(1/n)$

m balls are thrown into n bins.

X_i : number of balls in the i-th bin

union bound

$$\Pr\left[\max_{1\le i\le n} X_i \ge L\right] = \Pr\left[\exists 1 \le i \le n \text{ s.t. } X_i \ge L\right] \le n \Pr\left[X_i \ge L\right]$$

$$\Pr\left[X_i \ge L\right] \le \Pr\left[\exists L \text{ balls thrown into bin } i\right]$$

union bound

$$\le \sum_{S\in\binom{[m]}{L}} \Pr\left[\text{all balls in } S \text{ are thrown into bin } i\right]$$

$$= \binom{m}{L}\frac{1}{n^L} \le \frac{m^L}{L!n^L} \le \left(\frac{\mathrm{e}m}{n}\right)^L \frac{1}{L^L}$$

Stirling's approximation: $L! \ge \left(\frac{L}{\mathrm{e}}\right)^L$

m balls are thrown into n bins.

X_i : number of balls in the i-th bin

$$\Pr\left[\max_{1\le i\le n} X_i \ge L\right] \le n \Pr\left[X_i \ge L\right] \le \frac{1}{n}$$

$$\Pr\left[X_i \ge L\right] \le \left(\frac{\mathrm{e}m}{n}\right)^L \frac{1}{L^L}$$

$$\begin{cases} \text{(when } m = n\text{)} & \le \left(\dfrac{\mathrm{e}}{L}\right)^L \le \dfrac{1}{n^2} \qquad \text{for } L = \dfrac{3\ln n}{\ln\ln n} \\[2ex] \text{(when } m \ge n\ln n\text{)} & \le \left(\dfrac{L}{\mathrm{e}}\right)^L \dfrac{1}{L^L} = \mathrm{e}^{-L} \le \dfrac{1}{n^2} \end{cases}$$

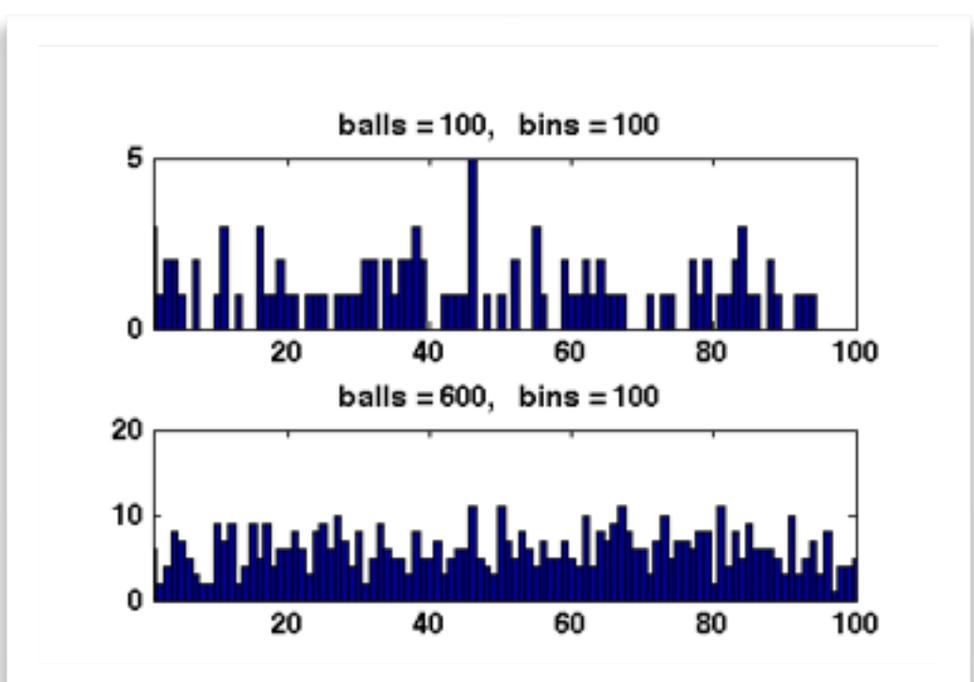
$$\text{for } L = \frac{\mathrm{e}^2 m}{n} \ge \mathrm{e}^2 \ln n$$

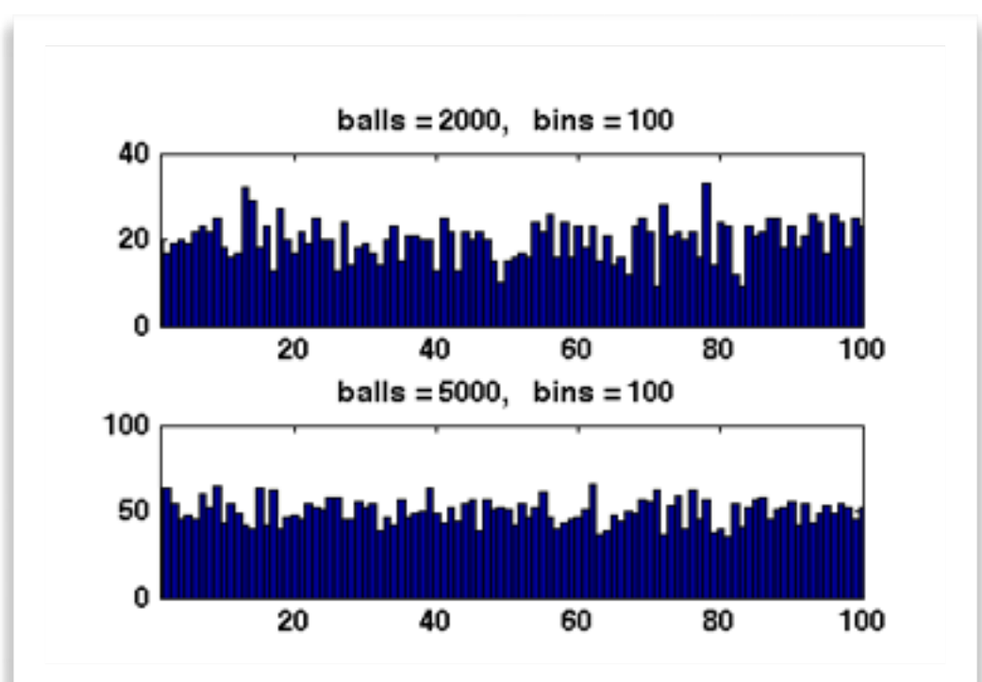
Occupancy Problem

- m balls are thrown into n bins uniformly and independently:

Theorem: With high probability, the maximum load is

$$\begin{cases} O\left(\dfrac{\log n}{\log\log n}\right) & \text{when } m = n \\ O\left(\dfrac{m}{n}\right) & \text{when } m \geq n \ln n \end{cases}$$



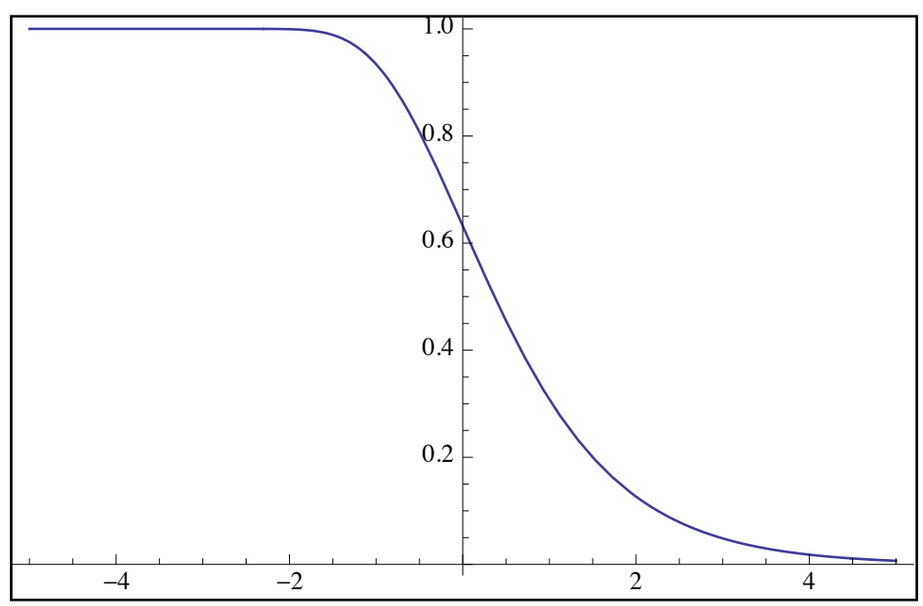


Coupon Collector

X : number of balls thrown to make all the n bins nonempty

a sharp threshold:

$$\lim_{n\to\infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$



Occupancy Problem

n balls are thrown into n bins.

X_i : number of balls in the i-th bin

Theorem:

With high probability, the maximum load is

$$\max_{1 \le i \le n} X_i = \Omega\left(\frac{\log n}{\log \log n}\right)$$

Poisson Heuristics

m balls are thrown into n bins.
X_i : number of balls in the i-th bin

- $X_1, \ldots, X_n$ are correlated binomial random variables:

$$X_1, \ldots, X_n \sim \text{Bin}(m, 1/n) \text{ subject to } \sum_{i=1}^{n} X_i = m$$

- *i.i.d.* Poisson random variables $Y_1, \ldots, Y_n \sim \text{Pois}(m/n)$

Poisson random variable $Y \sim \text{Pois}(\lambda)$:

$$\Pr[Y = k] = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0,1,2,\ldots$$

Poisson Heuristics

m balls are thrown into n bins.
X_i : number of balls in the i-th bin

- ***Heuristics***: treat loads of bins as *i.i.d.* $Y_1, \ldots, Y_n \sim$ Pois(m/n)

Poisson random variable $Y \sim$ Pois(λ):

$$\Pr[Y = k] = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \text{ for } k = 0,1,2,\ldots$$

(when $m = n \ln n + cn$)

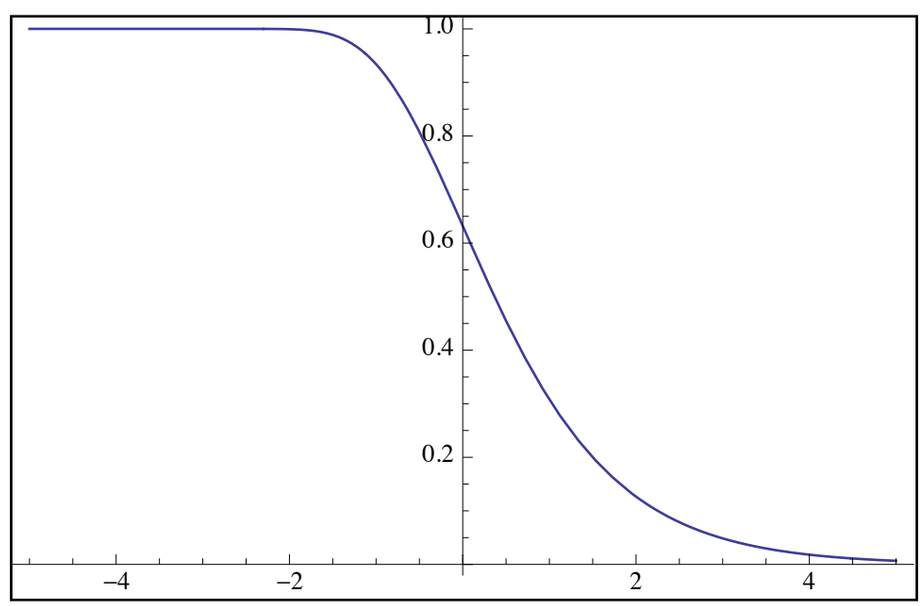
Coupon collector: $$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \left(1 - \mathrm{e}^{-\frac{m}{n}}\right)^n = \left(1 - \frac{\mathrm{e}^{-c}}{n}\right)^n \to \mathrm{e}^{-\mathrm{e}^{-c}}$$

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

a sharp threshold:

$$\lim_{n\to\infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$



Poisson Heuristics

m balls are thrown into n bins.
X_i : number of balls in the i-th bin

- ***Heuristics***: treat loads of bins as *i.i.d.* $Y_1, \ldots, Y_n \sim$ Pois(m/n)

Poisson random variable $Y \sim$ Pois(λ):

$$\Pr[Y = k] = \frac{\mathrm{e}^{-\lambda}\lambda^k}{k!} \text{ for } k = 0,1,2,\ldots$$

Occupancy problem:

$$\Pr\left[\max_{1\le i\le n} Y_i < L\right] = \left(\Pr[Y_i < L]\right)^n \le \left(1 - \Pr[Y_i = L]\right)^n$$

(when $m = n$) $$= \left(1 - \frac{1}{\mathrm{e}L!}\right)^n \le \mathrm{e}^{-n/(\mathrm{e}L!)} \le \frac{1}{n^2} \qquad \text{for } L = \frac{\ln n}{\ln\ln n}$$

since $L! \le \mathrm{e}\sqrt{L}\,(L/\mathrm{e})^L$

Occupancy Problem

n balls are thrown into n bins.

X_i : number of balls in the i-th bin

Theorem:

With high probability, the maximum load is

$$\max_{1 \le i \le n} X_i = \Omega\left(\frac{\log n}{\log \log n}\right)$$

Poisson Approximation

- loads of n bins receiving m balls: $X_1, \ldots, X_n$
- *i.i.d.* Poisson random variables $Y_1, \ldots, Y_n \sim \text{Pois}(\lambda)$

Theorem: $\forall m_1, \ldots . m_n \in \mathbb{N}$ s.t. $\sum_{i=1}^{n} m_i = m$

$$\Pr\left[\bigwedge_{i=1}^{n} X_i = m_i\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i = m_i \,\middle|\, \sum_{i=1}^{n} Y_i = m\right]$$

$$\Pr\left[\bigwedge_{i=1}^{n} X_i = m_i\right] = \frac{\binom{m}{m_1, \ldots, m_n}}{n^m} = \frac{m!}{m_1! \cdots m_n! n^m}$$

multinomial coefficient

$$\Pr\left[\bigwedge_{i=n}^{n} Y_i = m_i \,\middle|\, \sum_{i=1}^{n} Y_i = m\right] = \frac{\Pr\left[\bigwedge_i Y_i = m_i\right]}{\Pr\left[\sum_i Y_i = m\right]} = \frac{\prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{m_i}}{m_i!}}{e^{-n\lambda} \frac{(n\lambda)^m}{m!}} = \frac{m!}{m_1! \cdots m_n! n^m}$$

$$m = n \ln n + cn$$

Thm: *i.i.d.* $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ and $Y = \sum_{i=1}^{n} Y_i$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1)$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \sum_{k=0}^{\infty} \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = k\right] \Pr[Y = k]$$

choose

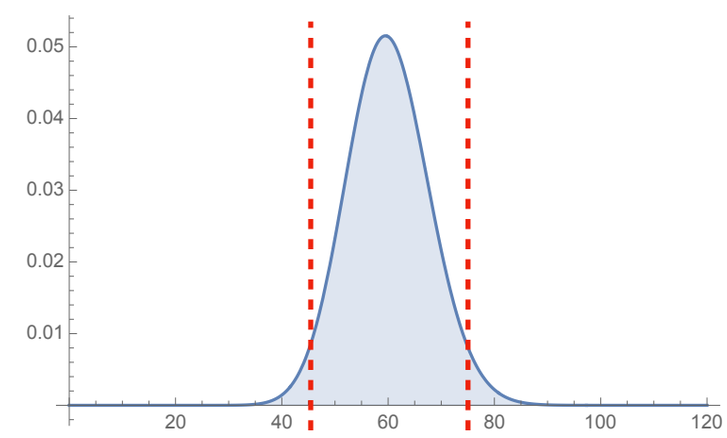
$$t = \sqrt{2m \ln m} \qquad \leq \sum_{k=m-t}^{m+t} \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = k\right] \Pr[Y = k]$$

$$+ \Pr[Y < m - t] + \Pr[Y > m + t] = o(1)$$

Lemma: $Y \sim \mathrm{Pois}(m)$

$$k > m \Rightarrow \Pr[Y > k] < \mathrm{e}^{-m} \left(\frac{\mathrm{e}m}{k}\right)^k$$

$$k < m \Rightarrow \Pr[Y < k] < \mathrm{e}^{-m} \left(\frac{\mathrm{e}m}{k}\right)^k$$



$$m = n \ln n + cn$$

Thm: *i.i.d.* $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ and $Y = \sum_{i=1}^{n} Y_i$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1)$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \sum_{k=m-t}^{m+t} \boxed{\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = k\right]} \Pr[Y = k] \; + o(1)$$

choose $t = \sqrt{2m \ln m}$

is monotone in k

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m - t\right] \le \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \le \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m + t\right]$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m + t\right] - \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m - t\right] \; \le ?$$

i.i.d. $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ conditioning on $Y = \sum_{i=1}^{n} Y_i = k$ is identically distributed as loads $X_1, \ldots, X_n$ for k balls into n bins

$$m = n \ln n + cn$$

Thm: *i.i.d.* $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ and $Y = \sum_{i=1}^{n} Y_i$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1)$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \sum_{k=m-t}^{m+t} \boxed{\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = k\right]} \Pr[Y = k] \; + o(1)$$

is monotone in k

choose $t = \sqrt{2m \ln m}$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m - t\right] \leq \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \leq \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m + t\right]$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m + t\right] - \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m - t\right]$$

$$\leq \Pr[\, 2t \text{ balls hit an empty bin}\,] \leq \frac{2t}{n} = o(1)$$

$$m = n \ln n + cn$$

Thm: *i.i.d.* $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ **and** $Y = \sum_{i=1}^{n} Y_i$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1)$$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \sum_{k=m-t}^{m+t} \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = k\right] \Pr[Y = k] \; + o(1)$$

$$= (1 - o(1)) \left(\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1) \right) + o(1)$$

$$m = n \ln n + cn$$

Thm: *i.i.d.* $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\frac{m}{n})$ and $Y = \sum_{i=1}^{n} Y_i$

$$\Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] = \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \pm o(1)$$

- loads of n bins receiving m balls: $X_1, \ldots, X_n$
- $(X_1, \ldots, X_n)$ is identically distributed as $(Y_1, \ldots, Y_n \mid Y = m)$

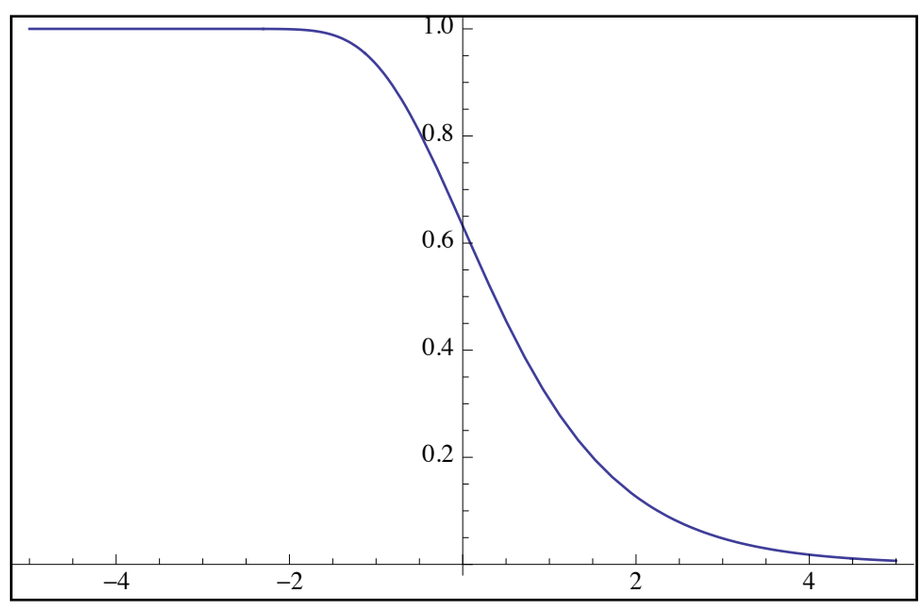
$$\begin{aligned}\Pr\left[\bigwedge_{i=1}^{n} X_i > 0\right] &= \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0 \mid Y = m\right] \\ &= \Pr\left[\bigwedge_{i=1}^{n} Y_i > 0\right] \pm o(1) \\ &\to \mathrm{e}^{-\mathrm{e}^{-c}} \quad \text{as } n \to \infty\end{aligned}$$

Coupon Collector

X : number of balls thrown to make all the n bins nonempty

a sharp threshold:

$$\lim_{n\to\infty} \Pr[X \geq n \ln n + cn] = 1 - \mathrm{e}^{-\mathrm{e}^{-c}}$$



Poisson Approximation

- loads of n bins receiving m balls: $X_1, \ldots, X_n$
- *i.i.d.* Poisson random variables $Y_1, \ldots, Y_n \sim \text{Pois}(m/n)$

Theorem (Poisson Approximation): $\forall$ nonnegative function f

$$\mathbb{E}\left[f(X_1, \ldots, X_n)\right] \leq \mathrm{e}\sqrt{m} \cdot \mathbb{E}\left[f(Y_1, \ldots, Y_n)\right]$$

$$\mathbb{E}\left[f(\overrightarrow{Y})\right] = \sum_{k=0}^{\infty} \mathbb{E}\left[f(\overrightarrow{Y}) \middle| Y = k\right] \Pr[Y = k]$$

where $Y = \sum_{i=1}^{n} Y_i \sim \text{Pois}(m)$

$$\geq \mathbb{E}\left[f(\overrightarrow{Y}) \middle| Y = m\right] \Pr[Y = m] = \mathbb{E}\left[f(\overrightarrow{X})\right] \mathrm{e}^{-m} \frac{m^m}{m!} \geq \frac{1}{\mathrm{e}\sqrt{m}} \mathbb{E}\left[f(\overrightarrow{X})\right]$$

since $m! \leq \mathrm{e}\sqrt{m}\left(\frac{m}{\mathrm{e}}\right)^m$

Poisson Approximation

- loads of n bins receiving m balls: $X_1, \ldots, X_n$
- *i.i.d.* Poisson random variables $Y_1, \ldots, Y_n \sim \text{Pois}(m/n)$

Theorem (Poisson Approximation): $\forall$ nonnegative function f

$$\mathbb{E}\left[f(X_1, \ldots, X_n)\right] \leq \mathrm{e}\sqrt{m} \cdot \mathbb{E}\left[f(Y_1, \ldots, Y_n)\right]$$

Occupancy problem:

$$\Pr\left[\max_{1 \leq i \leq n} X_i < L\right] \leq \mathrm{e}\sqrt{m}\,\Pr\left[\max_{1 \leq i \leq n} Y_i < L\right] = \mathrm{e}\sqrt{m}\left(\Pr[Y_i < L]\right)^n$$

(when $m = n$) $$\leq \mathrm{e}\sqrt{n}\left(1 - \Pr[Y_i = L]\right)^n = \mathrm{e}\sqrt{n}\left(1 - \frac{1}{\mathrm{e}L!}\right)^n$$

$$\leq \frac{1}{n} \qquad \text{for } L = \frac{\ln n}{\ln \ln n}$$

Occupancy Problem

n balls are thrown into n bins.

X_i : number of balls in the i-th bin

Theorem:

With high probability, the maximum load is

$$\max_{1 \le i \le n} X_i = \Omega\left(\frac{\log n}{\log \log n}\right)$$

Power of Two Choices

- n balls are thrown into n bins uniformly and independently
- with high probability, the maximum load is $\Theta\left(\frac{\log n}{\log\log n}\right)$

Two-Choice Paradigm:

balls are thrown into balls ***sequentially***;

for each ball:

uniformly and independently pick 2 bins; (with replacement)

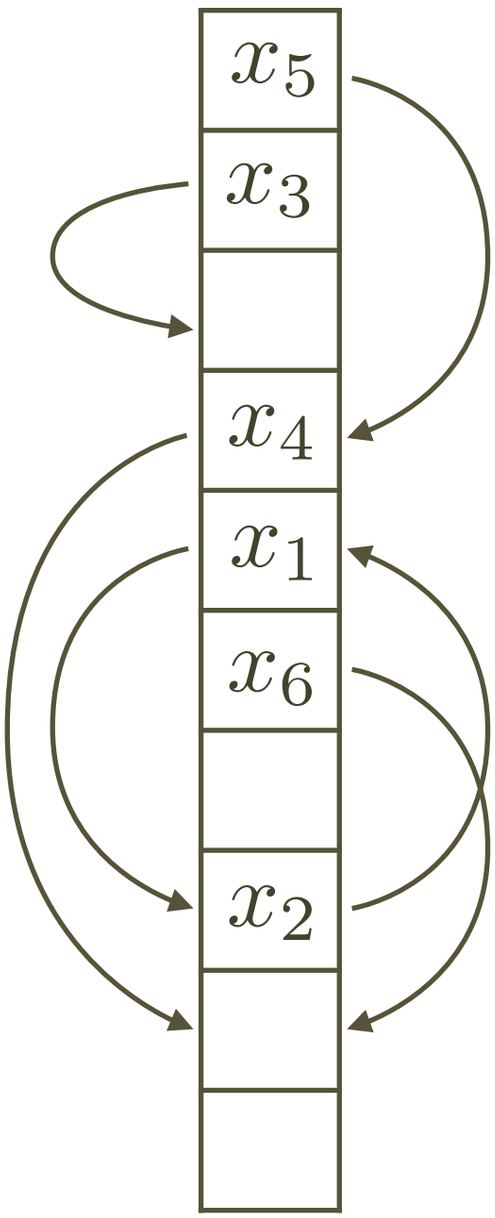
put the ball into the least full picked bin; (breaking tie randomly)

Theorem (Azar, Broder, Karlin, Upfal, 1999):

When n balls are thrown into n bins using the two-choice paradigm, the maximum load is $\ln\ln n/\ln 2 + \Theta(1)$ w.h.p.

Cuckoo Hashing

(Pagh 2001)



Cuckoo Hashing

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

uniform random $\boxed{h_1}$ $\boxed{h_2}$ $: [N] \to [m]$

m

Table T:

e	b	b	e d	c c	f		c c	f f	

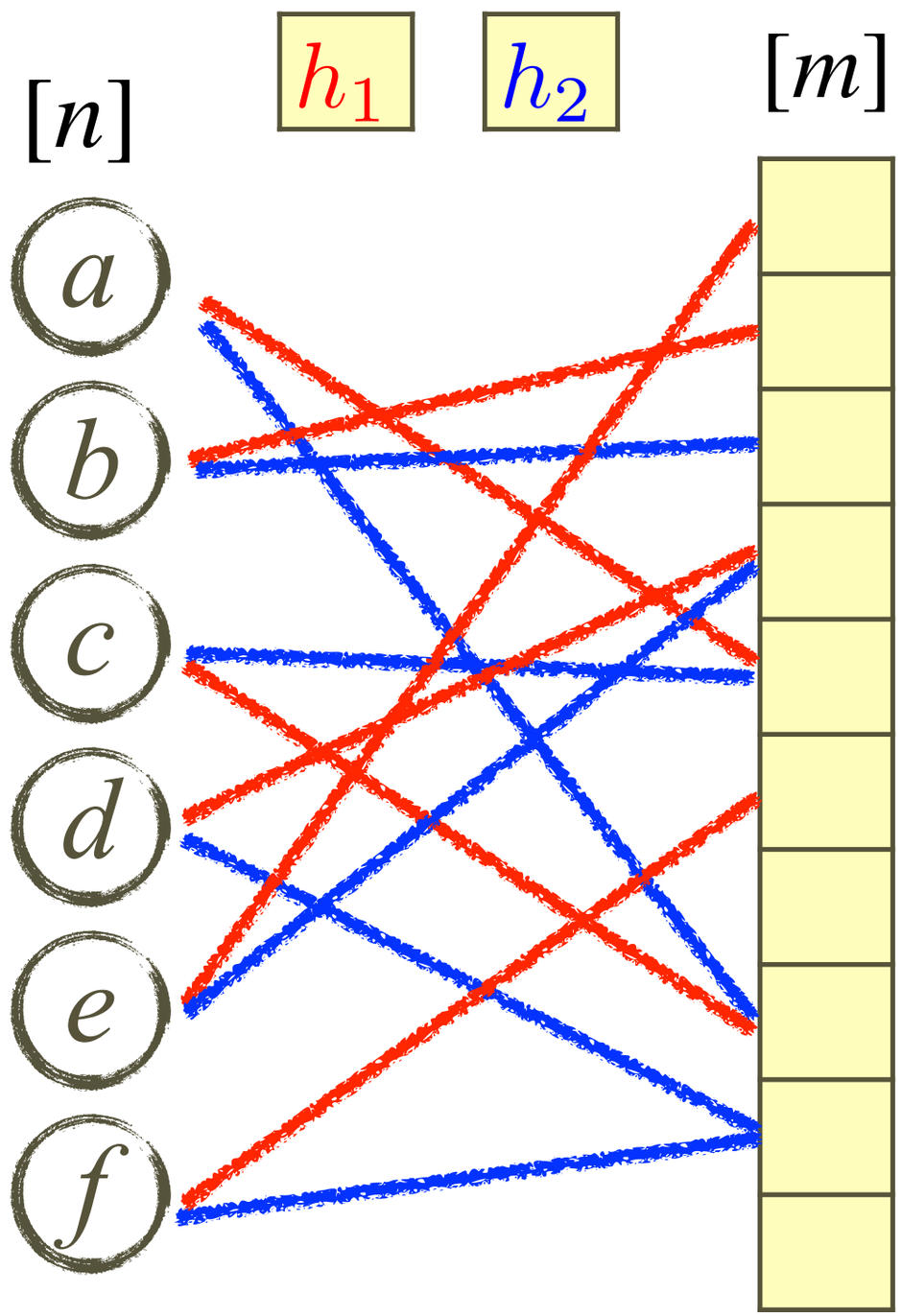
SUHA: Simple Uniform Hash Assumption

Query(x):

retrieve hash functions h_1 and h_2;

check whether $x \in \{T[h_1(x)], T[h_2(x)]\}$;

Cuckoo Hashing



Find such an $M : [n] \xrightarrow{1\text{-}1} [m]$ that:

$\forall x \in [n]$:

$$M(x) \in \{h_1(x), h_2(x)\}$$

M gives a matching of $[n]$ in bipartite graph $G([n], [m], E)$

$$(i, j) \in E \iff h_1(i) = j \vee h_2(i) = j$$

SUHA: uniform random $h_1, h_2 : [n] \to [m]$

$\Pr[\exists M] > 1/2$ for some $m = O(n)$