

**RECAP (KARGER - STEIN ALGORITHM)**

**OBSERVATION:** Full-contraction: early contractions much more valuable than later contractions in preserving min-cut  $C^*$

**IDEA:** "Share" valuable contractions  
**RECURSIVE STRATEGY**

$\text{R-KS}(G = (V, E))$ ;  $n \in |V|$

if ( $n \leq 8$ ) then \* solve directly \*

$$K \in n - \lceil n/\sqrt{2} + 1 \rceil \quad \{ K \approx (1 - \frac{1}{\sqrt{2}})n \approx n/3.4 \}$$

$$G_1 = (V_1, E_1) \in \text{PC}(G, K); t_1 \in \text{R-KS}(G_1) \{ T1 \}$$

$$G_2 = (V_2, E_2) \in \text{PC}(G, K); t_2 \in \text{R-KS}(G_2) \{ T2 \}$$

{ two independent trials,  $T1$  and  $T2$  }

$$\{ |V_1|, |V_2| \approx n/\sqrt{2} \approx n/1.41 \quad (> n/2) \}$$

return  $\text{MIN}(t_1, t_2)$

$$T(n) = \Theta(n^2 \log n) \quad T(n) = 2T\left(\frac{n}{\sqrt{2}}\right) + O(n^2)$$

**OBSERVATION:**  $h \leq \lceil 2 \log n \rceil + 1$  levels :  $2^h = \Theta(n^2)$   
 leaves! We perform  $\Theta(n^2)$  "full" contractions at the price of  $\log n$ ! But "full"

CONTRACTIONS ARE NOT INDEPENDENT ANYMORE (THEY SHARE COMMON SEGMENTS OF PARTIAL CONTRACTIONS)

# CORRECTNESS ANALYSIS

Let:  $p'(n) = \Pr(\text{f-ks}(g) \text{ with } n \text{ recursion steps is correct})$

For  $|N|=n : n \leq 2\lceil \log_2 n \rceil + 1$

We prove:  $p'(n) \geq \begin{cases} 1 & n=1 \\ p'(n-1) - \frac{1}{4}(p'(n-1))^2 & n>1 \end{cases}$

LEMMA

$$p'(n) \geq \frac{1}{n}$$

B:  $n=1$  OK  $1 \geq 1$  ( $p'(1)=1!$ )

HP:  $p'(n-1) \geq \frac{1}{n-1}, n \geq 2$

$$p'(n) \geq p'(n-1) - \frac{1}{4}(p'(n-1))^2 = f(p'(n-1)):$$

$$f(x) = x - \frac{1}{4}x^2: f'(x) = 1 - \frac{x}{2} > 0 \text{ in } [0, 1]$$

$f(x)$  increasing in  $[0, 1]$

Since  $p'(n-1) \geq \frac{1}{n-1}$  from HP

then

$$p'(n) \geq f(p'(n-1)) \geq f\left(\frac{1}{n-1}\right)$$

$$= \frac{1}{n-1} - \frac{1}{4(n-1)^2} > \frac{1}{n-1} - \frac{1}{2(n-1)^2} =$$

$$= \frac{1}{n-1} - \frac{1}{2(n-1)(n-1)} \geq \frac{1}{n-1} - \frac{1}{n(n-1)}$$

(since  $2(h-1) \geq h$  for  $h \geq 2$ )

$$= \frac{h-1}{h(h-1)} = \frac{1}{h}$$

Q.E.D.

We know that the number of levels of the recursion tree is  $h \leq \lceil 2 \log_2 n + 1 \rceil$

Therefore

$$\begin{aligned} p(u) &= \Pr(R\text{-KS}(G=(V,E)) \text{ returns } e^*, |V|=n) = p'(h) \geq \\ &\geq \frac{1}{h} \geq \frac{1}{\lceil 2 \log_2 n + 1 \rceil} \geq \frac{1}{4 \log_2 n} \end{aligned}$$

This is \*not\* high probability!

However we can AMPLIFY this probability

to multiple runs of R-KS:

KARGER-STEIN ( $G, S$ )

min  $\leftarrow +\infty$

repeat  $S$  times

$t \leftarrow R\text{-KS}(G)$

if ( $t < \text{min}$ ) then  $\text{min} \leftarrow t$

return  $\text{min}$

We have that

$\Pr(\text{KS}(G, S) \text{ fails returning } e^*) \leq$

$$\leq (1 - p(u))^S \leq \left(1 - \frac{1}{4 \log_2 n}\right)^S$$

Setting  $S = 4 \cdot d \cdot \log_2 n \ln n = \tilde{\Omega}(\log^2 n)$  we obtain:

$\Pr(\text{KS}(g, S) \text{ fails returning } 1e^* 1) \leq$

$$\leq \left[ 1 - \frac{\delta}{4 \log n} \right]^{4 \log n} \leq e^{-\delta} \leq e^{-1}$$
$$\leq e^{-\delta \log n} = \frac{1}{n^\delta}$$

$\Rightarrow \Theta(\log^2 n)$  executions of  $\text{RKs}(g)$  guarantee high probability. Therefore:

$$T_{\text{RKs}}(n) = O(n^2 \log n \cdot \log^2 n) = O(n^2 \log^3 n)$$

$$T_{\text{RKs}}(n) \quad \uparrow \quad \# \text{executions}$$

This is much faster than any MAX-FLOW based algorithm

### STATE OF THE ART :

- Deterministic direct algorithms (evading max-flow):  $O(n \log n \log \log n)$  (TOTALLY UNPRACTICAL, only theoretical)
- Karger's improvement over KS  $O(n \log^3 n)$  w.h.p.
- Many heuristic and approximation algorithms

## FURTHER RANDOMIZED TECHNIQUES:

### COUPON COLLECTING / RANDOMIZED ENUMERATION

Problem: Assume that an algorithm makes multiple calls to  $\text{RANDOM}(1,..,n)$ . How many calls are necessary to generate all numbers in  $\{1,..,n\}$ ?

Important primitive:

- How many card packets to complete collection of player cards?
- Collaborative work (e.g. SET@HOME) volunteers donate CPU time. Computational task divided into units. Each volunteer picks random unit. How many volunteer slots to execute all units?  
(Pros: decentralized, fault-tolerant  
Cons: redundancy )

### ANALYSIS IN EXPECTATION

Define  $Z_i = \#\text{ calls to } \text{RANDOM}(1,..,n)$  needed to obtain  $i$ -th distinct number

ATTENTION:  $i$ -th distinct number means that we have already extracted  $i-1$  distinct ones!

Then, of  $Z = \# \text{ of total cells}$ :

$$Z = \sum_{i=1}^m Z_i$$

OBSERVATION: We expect that, on average, the values of the  $Z_i$ 's increase

e.g.,  $\Pr(Z_1 = 1) = 1$  (the first cell always return a new value)

$$\Pr(Z_2 = 1) = 1 - \frac{1}{n} \quad (\text{very high})$$

$$\Rightarrow E[Z_2] = 1/(1-n) > 1, \text{ etc...}$$

In general: if we have already generated  $i-1$  distinct values, then each successive cell will return a new value with probability

$$p_i = \frac{n-(i-1)}{n} = \frac{n-i+1}{n}$$

$Z_i$  must then count the number of trials until the first head of a coin-flip with  $\Pr(\text{head}) = p_i$ .

$\Rightarrow Z_i \sim \text{Geom}(p_i)$ :

$$\Pr(Z_i = k) = p_i (1-p_i)^{k-1} \quad k \geq 1$$

We have

$$\begin{aligned} E[Z_i] &= \frac{1}{p_i} \quad (\text{exercise}) \\ &= \frac{n}{n-i+1} \end{aligned}$$

Therefore

$$\begin{aligned} E[z] &= E\left[\sum_{i=1}^n z_i\right] = \sum_{i=1}^n \frac{n}{n-i+1} \\ &= n \sum_{i=1}^n \frac{1}{n-i+1} = n \sum_{j=n-i+1}^n \frac{1}{j} \\ &= n H(n) = n (\ln n + O(1)) \\ &= n \ln n + O(n) \end{aligned}$$

## ANALYSIS IN HIGH PROBABILITY

We follow a different approach. Let us fix a given  $i \in \{1, \dots, n\}$ , and consider

$P_i = \Pr(\text{the first } r \text{ cells of } \text{RANDOM}(1, \dots, n) \text{ have not returned } i) = \Pr(E_i)$

OBSERVATION By symmetry:  $P_i = P_j = P(r, n)$

$\forall 1 \leq i \neq j \leq n$

We have

$$P(r, n) = \left(1 - \frac{1}{n}\right)^r = \left[\left(1 - \frac{1}{n}\right)^n\right]^{\frac{r}{n}} < e^{-r/n}$$

probability of returning a value  $\neq i$

Observe that by setting  $r = (d+1)n \ln n$

$$\text{we obtain } P(r, n) < \frac{1}{n^{d+1}}$$

$$r = (d+1)n \ln n$$

To finish the analysis:

$$\begin{aligned} & \Pr(\text{"all numbers observed after } (d+1) \text{ulum cells"}) \\ &= 1 - \Pr(\text{"at least one value is not observed after } (d+1) \text{ulum cells"}) \\ &= 1 - \Pr\left(\bigcup_{i=1}^m E_i\right) \stackrel{\text{union bound}}{\geq} 1 - \sum_{i=1}^m p_i = 1 - n \cdot P(C, n) \\ &\quad > 1 - \frac{1}{n^d} \end{aligned}$$

HIGH  
PROBABILITY!

# FURTHER RANDOMIZED TECHNIQUES:

## OCCUPANCY PROBLEMS / BALLS INTO BINS

EXPERIMENT:  $m \geq n$  identical balls are thrown randomly and independently into  $n$  bins (boxes)

AIM: Study events related to the final distribution of the balls into the bins.

### INTERESTING PHENOMENA

- 1) Maximum number of balls into any bin
- 2) Number of empty bins
- 3) How large must  $m$  be so that each bin contains at least a ball?

### SOME APPLICATIONS:

1. Analysis of hash functions under chaining: max number of collisions = max length of the list (max search time)
2. Idle workers under random distribution of load

3. Coupon collecting symmetry:  
 Balls = random extractions with replacement  
 Bins = values in the urn

3: We already know that  
 $m = \Theta(n \ln n)$  balls suffice to fill all bins w.h.p.  
 Also, on average,  $n H(n) = n \ln n + O(n)$  balls are necessary

1. MAXIMUM NUMBER OF BALLS INTO ANY BIN (after throwing  $m$  balls into  $n$  bins)

Let  $X_b = \text{"\# balls into bin } b\text{"}$ ,  $1 \leq b \leq n$   
 Clearly  $X_b \sim \text{Binomial}(m, \frac{1}{n})$   
 (sum of  $m$  Bernoulli trials)

$Y_j^b : 1 \leq j \leq m$

$$\Pr(Y_j^b = 1) = \Pr(\text{j-th ball into bin } b) = \frac{1}{n}$$

Thus:  $\Pr(X_b = i) = \binom{m}{i} \frac{1}{n^i} \left(1 - \frac{1}{n}\right)^{m-i}$

FACT  $\binom{m}{i} \leq \left(\frac{em}{i}\right)^i$  if  $m \geq 1$

STIRLING'S APPROXIMATION TO BIN. COEFF.

Let  $C_b(k)$  = "Bin b has  $\geq k$  balls"

Then:

$$\begin{aligned} \Pr(C_b(k)) &= \sum_{i=k}^m \binom{m}{i} \left(\frac{1}{m}\right)^i \left(1 - \frac{1}{m}\right)^{m-i} \\ &\leq \sum_{i=k}^m \left(\frac{em}{i}\right)^i \left(\frac{1}{m}\right)^i = \sum_{i=k}^m \left(\frac{em}{im}\right)^i \end{aligned}$$

for  $i \geq k$ :

$$\left(\frac{em}{im}\right)^i \leq \left(\frac{em}{km}\right)^i$$

thus

$$\begin{aligned} \Pr(C_b(k)) &\leq \sum_{i=k}^m \left(\frac{em}{km}\right)^i \leq \sum_{i=k}^{\infty} \left(\frac{em}{km}\right)^i \\ &= \left(\frac{em}{km}\right)^k \sum_{j=0}^{\infty} \left(\frac{em}{km}\right)^j \\ &= \left(\frac{em}{km}\right)^k \frac{1}{1 - \frac{em}{km}} \stackrel{\text{HP: } (k \geq 2em/m)}{\approx} 2 \left(\frac{em}{km}\right)^k \end{aligned}$$

$\hookrightarrow 1.49 \text{ da dimensio} \dots$

Consider the case  $m = n$

Letting  $k^* = \lceil (3\ln n) / \ln \ln n \rceil$  we obtain:

$$\Pr(C_b(k^*)) \leq 2 \left( \frac{e}{k^*} \right)^{k^*} = 2(e^{1-\ln k^*})^{k^*}$$

$$\leq 2 \left( e^{\underbrace{(1-\ln 3 - \ln \ln n + \ln \ln \ln n)}_{\ln k^*}} \right)^{k^*}$$

$$< 2 \cdot e^{(-\ln \ln n + \ln \ln \ln n)^{k^*}}$$

$$\leq 2 \cdot e^{-3\ln n + 4\ln n} \frac{\ln \ln \ln n}{\ln \ln n} \xrightarrow{n \rightarrow 0}$$

$$< \frac{1}{8}$$

$$\leq 2 \cdot e^{-\frac{3\ln n + \ln n}{2}} =$$

$$n > \bar{n}$$

$$\approx 2 \cdot e^{-2.5\ln n} = 2 \cdot \frac{1}{n^{2.5}} \leq \frac{1}{m^2} (\bar{n}^2)$$

Thus

$$\Pr(\exists b \in [m] \text{ with } \geq k^* \text{ balls}) = \Pr\left(\bigcup_{b=1}^m C_b(k^*)\right)$$

UNION BOUND

$$\leq \sum_{b=1}^m \Pr(C_b(k^*)) \leq \frac{m}{m^2} = \frac{1}{m}$$

Therefore the maximum number  
of balls into a bin when  $M = n$ ,  
is

$$O\left(\frac{\log n}{\log \log n}\right)$$

w.h.p.

### COROLLARY

Under a fully random hash  
function with chaining, the  
search time for a key is

$$O\left(\frac{\log n}{\log \log n}\right)$$

w.h.p.

( better than any deterministic  
dictionary tree structure )