

# Ausgewählte Kapitel ADS

December 16, 2015

## 1 Datenstrukturen für Mengen

### 1.1 Union-Find-Problem

Verwaltung von diskunkten Mengen

#### Problem

Verwalte eine Partition (Zerlegung in disjunkte Teilmengen) der Menge  $\{1, \dots, n\}$  unter folgenden Operationen.  
Jede Teilmenge (Block) besitzt einen eindeutigen Namen aus  $\{1, \dots, n\}$ .

- FIND( $x$ ):  $x \in \{1, \dots, n\}$  Liefert den Namen der Teilmenge, die  $x$  enthält
- UNION( $A, B, C$ ): Vereinigt die Teilmengen mit Namen  $A$  und  $B$  zu einer Teilmenge mit dem Namen  $C$ .

#### Initialisierung

Wir starten mit der Partitionierung:  $\{\{1\}, \dots, \{n\}\}$  mit dem Namen  $i$  für  $\{i\}, 1 \leq i \leq n$

Analyse: Kosten für 1 Union (worst case)

Amortisiert: Kosten für  $n - 1$  mögliche UNIONS

→ Kosten von  $n - 1$  UNIONs und  $m$  FINDs

#### Lösungen

##### 1. Lösung (einfach)

Verwende ein Feld  $\text{name}[1..n]$  mit  $\text{name}[x] = \text{Name des Blocks der } x$  enthält.  $1 \leq x \leq n$

```
for i=1 to n do
    name[ i ] <- i
od
FIND(x): return name[x] :  $\mathcal{O}(n)$ 
UNION(A,B,C):  $\mathcal{O}(n)$ 
for i=1 to n do
    if name[ i ] = A OR name[ i ] = B
        then name[ i ] <- C
    fi
od
```

Gesamlaufzeit (Lemma 1):

$n - 1$  UNIONs und  $m$  FINDs kosten  $\mathcal{O}(n^2 + m)$

##### 2. Lösung (Verbesserung)

1. Find unverändert

2. Ändere den Namen der kleineren Menge in den Namen der größeren (Relabel the smaller half)

Zusätzliche Felder:

- $\text{size}[1..n]$ :  $\text{size}[A] = \text{Anzahl Elemente im Block } A$ , initialisiert mit 1
- $L[1..n]$ :  $L[A] = \text{Liste aller Elemente in Block } A$ , initialisiert  $L[i] = \{i\}$

$\text{FIND}(x)$  bleibt gleich

$\text{UNION}(A,B)$ :

```

if size [A] ≤ size [B]
then
    forall i in L[A] do
        name[i] ← B
    od
    size[B] += size[A]
    L[B] ← L[B] concatenate L[C]
else
    symmetrisch

```

Die Menge heißt jetzt A oder B

Effekt:  $\text{UNION}(A,B,..)$  hat Laufzeit  $\mathcal{O}(\min(|A|, |B|))$

Worst Case eines UNION dieser Folge von UNIONS:  $\mathcal{O}\left(\frac{n}{2}\right) = \mathcal{O}(n)$  (kann nur einmal vorkommen)

Wie oft kann sich  $\text{name}[x]$  für ein bestimmtes  $x : 1 \leq x \leq n$  ändern?

Beobachtung:

- Am Anfang ist jedes Element  $x$  in einer ein-elementigen Menge
- Am Ende sind alle Elemente in einer Menge der Größe  $n$
- Immer wenn ein Element  $x$  seinen Namen ändert befindet es sich danach in einer doppelt so großen Menge (nach dem UNION)

⇒ Jedes Element  $x \in \{1, \dots, n\}$  kann maximal  $\log(n)$  mal seinen Namen ändern.

Satz 1: Bei UNION-FIND mit "Relabel the smaller half" sind die Gesamtkosten einer beliebigen Folge von  $n-1$  UNIONS und  $m$  Finds  $\mathcal{O}(m + n * \log(n))$

Im Schnitt (amortisiert) kostet ein UNION  $\log(n)$

### 3. Lösung

Lösung 1 und 2 haben FIND effizient gelöst, hier UNION

Jeder Block wird als Baum dargestellt. Die Knoten repräsentieren die Elemente des Blocks. In der Wurzel steht der Name des Blocks.

$\text{UNION}(A,B,E)$ : Mache die Wurzel von A zum Kind der Wurzel von B und nenne die Wurzel um in E.

$\text{FIND}(x)$ : Starte bei Element (Knoten)  $x$  und laufe bis zur Wurzel, dort steht der Name →  $\mathcal{O}(\text{Tiefe von } x)$

#### Realisierung der Datenstruktur durch Felder:

$$\text{vater}[i] = \begin{cases} \text{Vater von } i \text{ in seinem Baum} \\ 0, \text{ falls } i \text{ Wurzel} \end{cases}$$

$\text{name}[i] = \text{Name des Blocks mit Wurzel } i$  (at nur Bedeutung, falls  $i$  Wurzel)

$\text{wurzel}[i] = \text{Wurzel des Blocks mit Namen } i$

Initialisierung:

```

for i=1 to n do
    vater[i] = 0
    name[i] = i
    wurzel[i] = i
od

```

FIND(x):

```

while vater[x] != 0 do
    x = vater[x]
od
return name[x]

```

UNION(A,B,C):

```

r1 = wurzel[A]
r2 = wurzel[B]
vater[r1] = r2
name[r2] = C
wurzel[C] = r2

```

Analyse:

- UNION:  $\mathcal{O}(1)$  worst case
- FIND(x): Tiefe von  $x$  (max Höhe des entstehenden Baums,  $n-1$  möglich)

### 4. Lösung (Weighted Union rule):

Vermeide große Tiefen, dafür hänge den kleineren Baum (Anzahl Knoten) an den größeren

Alternativ: Hänge den Baum mit kleinerer Höhe an den tieferen.

Realisierung: Zusätzliches Feld

$\text{size}[i]$  = Anzahl Knoten um Unterbaum mit Wurzel i

Initialisierung:

FIND(x) (wie bei 3):

```
for i=1 to n do
    vater[i] = 0      while vater[x] != 0 do
        name[i] = i          x = vater[x]
        wurzel[i] = i      od
        size[i] = 1      return name[x]
od
```

Laufzeit  $\mathcal{O}(\log(n))$ :

Sei für jeden Knoten x die  $\text{höhe}(x)$  die Höhe von x in seinem Baum (maximale Pfad zu Blatt), Blatt=0  
 $\text{size}(x)$ : Anzahl der Knoten im Unterbaum mit Wurzel x (Gewicht)

Lemma: Bei weighted Union rule gilt stets, dass  $\text{size}(x) \geq 2^{\text{höhe}(x)}$  für alle Knoten x.

Beweis: Induktion über  $\text{höhe}(x)$ :

Voraussetzung:

$\text{höhe}(x) = 0$ : x ist Blatt  $\rightarrow \text{size}(x) = 1 = 2^0$

Anfang:

$\text{size}(y) \geq 2^{\text{höhe}(y)}$

Schritt:

Sei  $\text{höhe}(x) > 0$

Sei y ein Kind von x mit  $\text{höhe}(x)-1$

Betrachte die UNION Operation bei der x zum Vater von y wurde.

Seien  $\overline{\text{size}}(x)$  und  $\overline{\text{size}}(y)$  die Gewichte vor der UNION Operation, dann gilt:

1)  $\text{size}(y) = \overline{\text{size}}(y)$ , da sich das Gewicht nur für Wurzeln ändern kann

2)  $\overline{\text{size}}(x) \geq \text{size}(y)$  durch weighted union rule

3) Nach der Operation:  $\text{size}(x) \geq \overline{\text{size}}(x) + \overline{\text{size}}(y)$

$\geq 2 * \overline{\text{size}}(y)$  wegen 2.

$\geq 2 * \overline{\text{size}}(y)$  wegen 1.

$\geq 2 * 2^{\text{höhe}(y)}$  nach IA

$= 2^{\text{höhe}(y)+1} = 2^{\text{höhe}(x)}$

Da Anzahl der Knoten  $n \Rightarrow \text{size}(x) \leq n$  gilt:

$\Rightarrow n \geq \text{size}(x) \geq 2^{\text{höhe}(x)}$  für alle x

$\text{höhe}(x) \leq \log(n)$

Satz: Bei UNION-FIND mit weighted UNION ist die Laufzeit einer beliebigen Folge  $n-1$  Unions und  $m$  Finds  $\mathcal{O}(n + \log(n))$

Beweis: 1. UNION  $\mathcal{O}(1)$  worst-case, 2. Find  $\mathcal{O}(\log(n))$  worst case (Lemma)

### 5. Lösung (Verbesserung von FIND):

Pfad-Komprimierung (path compression)

Ein FIND(x) durchläuft den Pfad von x zur Wurzel.

$x = x_0, \dots, x_l = \text{Wurzel}$

Idee: Hänge  $x_0, \dots, x_{l-1}$  direkt an die Wurzel an.

Erhöht die Kosten dieses Finds um einen konstanten Faktor.

Algorithmus:

FIND(x)

```
r <- x;
while vater[r] != 0 do
    r <- Vater[r]
od
while x != r do
    y <- vater[x]
```

UNION(A,B,C):

```
r1 = wurzel[A]
r2 = wurzel[B]
if size[r1] ≤ size[r2] then
    vater[r1] = r2
    name[r2] = C
    wurzel[C] = r2
    size[r2] += size[r1]
else
    symmetrisch
```

```

vater [ x ] <- r
x <= y
od

```

Ganz klar:  $\mathcal{O}(\log(n))$  worst case

Satz (Tarjan): Bei UNION-FIND mit weighted UNION und path compression hat eine beliebige Folge von n-1 Unions und m Finds mit  $m \geq n$ , die Gesamtkosten  $\mathcal{O}(m * \alpha(m, n))$ , wobei  $\alpha(m, n) = \min\{z \in \mathbb{N} | A(z, \frac{4m}{n}) > \log n\}$

mit A einer Variante der Ackermannfunktion.

$\alpha$  ist eine Art Inverse der Ackermannfunktion  $\Rightarrow$  Ist extrem langsam wachsend.

Definition von A:

$$A : \mathbb{N}_0 \times \mathbb{N}_0 \rightarrow \mathbb{N}_0$$

$$A(i, 0) = 0 \text{ für alle } i \in \mathbb{N}_0$$

$$A(0, x) = 2x \text{ für alle } x \geq 1$$

$$A(i, 1) = 2$$

$$A(i, x) = A(i - 1, A(i, x - 1)) \text{ für } i \geq 1, x \geq 2$$

$$\begin{array}{cccccc} 0 & 2 & 4 & 6 & 8 & 10 \\ 0 & 2 & 4 & 8 & 16 & 32 \end{array}$$

$$\begin{array}{cccccc} A(i, x) \text{ als Matrix } i \times x: & 0 & 2 & 4 & 16 & 65536 & 2^{65536} \\ & 0 & 2 & 4 & 65536 & 2 \uparrow\uparrow 65536 & . \\ & 0 & 2 & . & . & . & . \end{array}$$

$$1. \text{ Zeile: } A(0, x) = 2x; 2. \text{ Zeile: } A(1, x) = 2^x; 3. \text{ Zeile: } A(2, x) = 2^{2^x}$$

Anmerkung: Pfeilschreibweise=(Knuth Up-Arrow)

Beweis des Satzes:

Situation: n Elemente {1,...,n}, beliebige Folge von n-1 Unions und m Finds:  $U_1, F_1, F_2, U_2, \dots$

Am Ende: 1 Baum T' (n-1 weighted Unions)

Konzeptuell kann T' anders erhalten werden: Führe zunächst alle Unions aus  $\rightarrow$  Baum T. Dann führe m partielle Finds auf T aus ( $PF_1, \dots, PF_m$ ), die genau den selben Pfad wir  $F_i$  durchlaufen, bis zu ihrer ursprünglichen Wurzel vor den Unions.

Wir schätzen nun die Gesamtkosten dieser Folge (insbesondere der m PF's) ab.

Frage: Wieviele Vater-Verweise (Kanten) werden insgesamt durchlaufen?

Sei F=Multi-Menge aller durch die PF's durchlaufenden Kanten (mit Mehrfachen)

Zu zeigen:  $|F| = \mathcal{O}(m * \alpha(m, n))$

Idee:

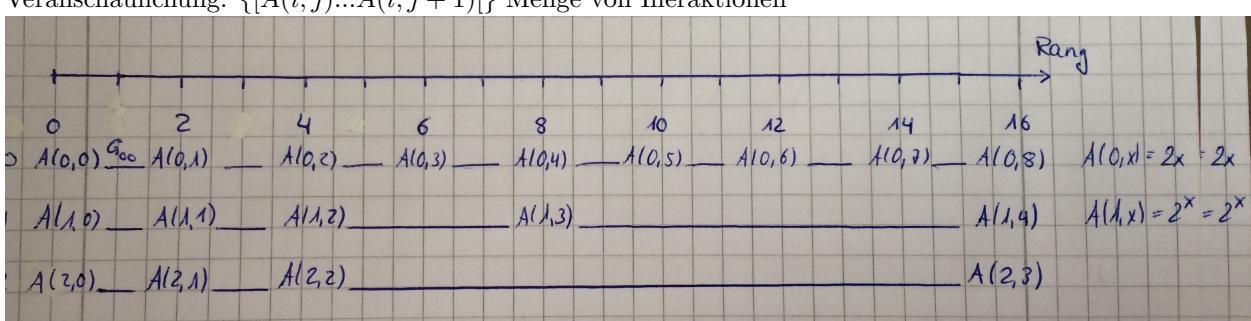
- Teile F in Gruppen nach Rang der Endpunkte der Kanten, wobei  $\text{Rang}(x)=\text{Höhe}(x)$  im Baum T' (nicht T)
- Schätzt die Gruppen einzeln ab.

Zunächst Einteilung der Knoten in die Gruppen nach Rang (nicht disjunkt).

Sei  $z \in \mathbb{N}_0$ , Für  $0 \leq i \leq z, j \geq 0$  sei:

$$G_{i,j} = \{Knoten x | A(i, j) \leq \text{Rang}(x) < A(i, j + 1)\}$$

Veranschaulichung:  $\{[A(i, j) \dots A(i, j + 1)]\}$  Menge von Interaktionen



Beispiel:  $\text{Rang}(x)=7$ ,  $\text{Rang}(y)=13$

$\Rightarrow x \in G_{0,3}, G_{1,2}, G_{2,2}, \dots$

$\Rightarrow y \in G_{0,6}, G_{1,3}, G_{2,2}, \dots$

Eine Einteilung der Multi-Menge F

Für  $0 \leq k \leq z$ :  $N_k = \{(x, y) \in F | k = \min\{i \geq 0 | \exists j \text{ mit } x, y \in G_{i,j}\}\}$

und  $N_{z+1} := F \setminus \bigcup_{0 \leq i \leq z} N_i$

Schließlich definieren wir für  $0 \leq k \leq z+1 : L_k = \{(x, y) \in N_k | (x, y) \text{ ist letzte (oberste) Kante auf PF-Pfad}\}$

- a)  $|L_k| \leq m$  für  $0 \leq k \leq z+1$
- b)  $|N_0 \setminus L_0| \leq n$
- c)  $|N_k \setminus L_k| \leq \frac{5}{8}n$  für  $1 \leq k \leq z$
- d)  $|N_{z+1} \setminus L_{z+1}| \leq n * a(z, n)$  mit  $a(z, n) = \min\{i \geq 0 | A(z, i) > \log(n)\}$

Beweis:

a) Für jedes PF gibt es höchstens 1 Kante in  $L_k$ . Die Behauptung folgt daraus, dass es insgesamt nur m PFs gibt.

b) Sei  $(x, y) \in N_0 \setminus L_0$ , dann gilt  $\exists j \geq 0$  mit  $x, y \in G_{0,j}$ , das heißt  $A(0, j) \leq Rang(x) < Rang(y) < A(0, j+1)$   
 $\Rightarrow Rang(x) = 2j, Rang(y) = 2j + 1$

$(x, y) \notin L_0 \Rightarrow$  nicht die letzte Kante in diesem PF: Betrachte PF von (x,y), dann existiert eine Kante  $(s, t) \in L_0$  auf diesem PF-Pfad

Situation:

$$Rang(x) = 2j, Rang(y) = 2j + 1$$

$Rang(s) \geq Rang(y)$  da letzter Pfad vor dem nicht letzten sein muss

$Rang(t) > Rang(s)$  da Pfad von s nach t

$$\Rightarrow Rang(t) \geq 2j + 2$$

Nach dem PF: x hat neuen Vater (möglicherweise t) u mit  $Rang(u) \geq Rang(t) \geq 2k + 2$

$$\Rightarrow \text{Rangdifferenz zwischen x und dem neuen Vater u} \geq 2$$

$\Rightarrow$  Spätere PFs können keine Kante (x,..) mehr zu  $N_0$  hinzufügen

$\Rightarrow$  Für jeden Knoten wird maximal eine ausgehende Kante (Vaterverweis) in  $N_0 \setminus L_0$  gezählt werden.

$$\Rightarrow |N_0 \setminus L_0| \leq n$$

Beweis c), d):

Idee: Schätze Beitrag eines Knotens  $x \in G_{k,j}$  zu  $N_k \setminus L_k$  d.h. alle Kanten, die von x ausgehen und in  $N_k \setminus L_k$  gezählt werden.

Sei  $k \geq 1$  und  $x \in G_{k,j}$  beliebig, d.h.  $\exists j$  mit  $A(k, j) \leq Rang(x) < A(k, j+1)$

und  $y_1, \dots, y_q$  alle Endknoten mit  $(x, y_i) \in N_k \setminus L_k$ . Ziel: q nach oben abschätzen.

$$\Rightarrow Rang(y_1) \leq \dots \leq Rang(y_q) < A(k, j+1)$$

Beobachtungen:

1)  $j \geq 2$  weil sonst k=0 die minimale Zeile definiert, sodass  $(x, y_i)$  im selben Intervall (die ersten 3 Spalten sind immer gleich gefüllt mit 0,2,4). Hier:  $k \geq 1$

2)  $(x, y_i) \notin L_k$  für  $1 \leq i \leq q \Rightarrow \exists (s_i, t_i) \in N_k$  auf PF-Pfad von  $(x, y_i)$  oberhalb von  $(x, y_i)$  Nach der Pfadkopplermierung ist Vater von  $x = y_i + 1$ . Außerdem ist  $y_{i+1}$  Vorfahr von  $t_i$  dabei ist  $t_i = y_{i+1}$  möglich.

Es gilt stets  $Rang(x) < Rang(y_i) \leq Rang(s_i) < Rang(t_i) \leq Rang(y_{i+1})$

Definition von  $N_k$  k minimal  $\Rightarrow (x, y_i), (s_i, t_i) \notin N_{k-1}$

$$\Rightarrow \exists j \text{ mit } Rang(s_i) < A(k-1, j) \leq Rang(t_i)$$

Daher gilt  $Rang(y_i) < A(k-1, j) \leq Rang(y_{i+1})$  für  $1 \leq i \leq q-1$

Anwendung auf die gesamte Folge  $y_1, \dots, y_q$  (d.h. q-1 mal):

$$Rang(y_1) < A(k-1, j_1) \leq Rang(y_2) < A(k-1, j_2) \leq Rang(y_3) < \dots \leq Rang(y_{q-1}) < A(k-1, j_{q-1}) < Rang(y_q)$$

Beobachtung:  $j_{i+1} \geq j_i \Rightarrow \exists j_1 \text{ mit } Rang(y_1) < A(k-1, j_1) \leq A(k-1, j_1 + q-1) \leq Rang(y_q)$

**I.**  $\exists j \geq 2 : Rang(y_q) \geq A(k-1, j + q - 1)$

Beweis Teil c):

$k \geq 1, x \in G_{k,j}, (x, y_i) \in N_k \Rightarrow y_1, \dots, y_q \in G_{k,j}, j \geq 2$

$$\Rightarrow A(k, j) \leq Rang(y_1) \leq \dots \leq Rang(y_q) < A(k, j+1)$$

**II.**  $Rang(y_q) < A(k, j+1)$

Nach I und II:

$$\exists j : A(k-1, j + q - 1) \leq A(k, j+1) = A(k-1, A(k, j))$$

$\Rightarrow$  (Monotonie von A in Zeilen)  $j + q - 1 < A(k, j)$

$$\Rightarrow (j \geq 2)q < A(k, j)$$

Wir haben gezeigt: Für jeden Knoten  $x \in G_{k,j}, k \geq 1, j \geq 2$  gibt es höchstens A(k,j) Kanten  $(x, y) \in N_k \setminus L_k$

$$\Rightarrow |N_k \setminus L_k| \leq \sum_{j \geq 2} |G_{k,j}| * A(k, j) \text{ mit } 1 \leq k \leq z$$

Behauptung:  $|G_{k,j}| \leq \frac{2n}{2^{A(k,j)}}$  extrem fallend, n Knoten im Baum.

Daraus folgt:  $|N_k \setminus L_k| \leq \sum_{j \geq 2} \frac{2n * A(k, j)}{2^{A(k, j)}}$  wobei  $A(k, j) \leq 2^j$ , da  $k \geq 1$  zweite Zeile.

$\leq 2n \sum_{j \geq 2} \frac{2^j}{2^{2^j}} = 2n \sum_{j \geq 2} \frac{1}{2^{2^j-j}} = 2n(\frac{1}{4} + \frac{1}{32} + \frac{1}{2^{12}} + \dots)$  wobei  $(\frac{1}{32} + \dots)$  mit  $\frac{1}{16}$  abgeschätzt wird.  
 $\Rightarrow = \frac{5}{8}n$

Beweis der Behauptung  $|G_{k,j}| \leq \frac{2n}{2^{A(k,j)}}$

Sei  $l$  beliebig mit  $A(k,j) \leq l < A(k,j+1)$

Zähle zuerst alle Knoten mit  $Rang(x) = l$ . Dafür sei  $G_{k,j,l} = \{x \in G_{k,j} \mid Rang(x) = l\}$

Es gilt:

1. Jeder Knoten  $x$  mit  $Rang(x) = l$  hat mindestens  $2^l$  Nachkommen (alle Knoten im Unterbaum mit Wurzel  $x$ ), nach Lemma weighted Union

2. Für  $x \neq y$  mit  $Rang(x) = Rang(y)$  sind die Nachkommensmengen disjunkt. 1+2:  $|G_{k,j,l}| \leq \frac{n}{2^l}$  mit  $n$  Gesamtanzahl der Knoten

$$\begin{aligned} \Rightarrow |G_{k,j}| &= \sum_{l=A(k,j)}^{A(k,j+1)-1} |G_{k,j,l}| \\ &\leq \sum_{l=A(k,j)}^{\inf} \frac{n}{2^l} = n \sum_{l=A(k,j)}^{\inf} \frac{1}{2^l} \\ &= n \left( \frac{1}{2^{A(k,j)}} + \frac{1}{2} \frac{1}{A(k,j)} + \frac{1}{4} \frac{1}{A(k,j)} + \dots \right) \\ &= n * \frac{2}{2^{A(k,j)}} \end{aligned}$$

Beweis Teil d:

$k = z + 1$ , weighted Union  $\Rightarrow Rang(y_q) \leq \log(n)$

Nach I. für  $k = z + 1$ :

$A(z, j + q - 1) \leq Rang(y_q) \leq \log(n)$

$\Rightarrow j + q - 1 < \alpha(z, n)$ , da  $\alpha(z, n)$  minimal mit  $A(z, \alpha(z, n)) > \log(n)$

$\Rightarrow q < \alpha(z, n)$  mit  $j \leq 2$

Also gibt es für jeden Knoten  $x$  höchstens  $\alpha(z, n)$  Kanten  $(x, \dots) \in N_{z+1} \setminus L_{z+1}$  mit  $n = \text{Anzahl aller Knoten}$

$\Rightarrow |N_{z+1} \setminus L_{z+1}| \leq n * \alpha(z, n)$

Beweis des Satzes (Tarjan):

Jede beliebige Folge von  $n-1$  Unions und  $m \geq n$  Finds hat die Gesamtaufzeit von  $\mathcal{O}(m * \alpha(m, n))$

Lemma (Kosten aller Finds): Für jedes  $z \geq 0$  gilt:

$$\begin{aligned} |F| &= \sum_{k=0}^{z+1} |L_k| + \sum_{k=1}^{z+1} |N_k \setminus L_k| = \sum_{k=0}^{z+1} |L_k| + |N_0 \setminus L_0| + \sum_{k=1}^z |N_k \setminus L_k| + |N_{z+1} \setminus L_{z+1}| \\ &\leq (z+2) * m + n + \frac{5}{8}n * z * + n * \alpha(z, n) \end{aligned}$$

Betrachte  $z = \alpha(m, n)$ ,  $n \leq m$  (jedes  $z$  liefert eine obere Schranke)

Dann gilt:

$$|F| \leq \mathcal{O}(m * \alpha(m, n)) + n + \mathcal{O}(n * \alpha(m, n)) + n * \alpha(z, n)$$

$$|F| \leq \mathcal{O}(m * \alpha(m, n) + n * \alpha(z, n))$$

$$\Rightarrow \alpha(\alpha(m, n), n) \leq 4 \frac{m}{n}$$

$$\Rightarrow \leq n * \frac{4m}{n} = \mathcal{O}(m)$$

Insgesamt: Kosten aller Find-Operationen sind  $|F| = \mathcal{O}(m\alpha(m, n) + m) = \mathcal{O}(m\alpha(m, n))$

Kosten aller Unions:  $\mathcal{O}(n) = \mathcal{O}(m)$

Bemerkung:

1) In der Praxis sehr gute Laufzeiten (sehr einfache Algorithmen und Datenstrukturen,  $\alpha(m, n) < 4$  für alle in der Praxis vorkommenden Werte von  $m, n$ ).

2) Die Schranke  $m\alpha(m, n)$  ist scharf, d.h. das Union-Find-Problem hat tatsächlich diese Komplexität. Es gibt Beweise für untere Schranke  $\Omega(m * \alpha(m, n))$

3) Optimal...yeay.

4) Varianten: Split-Find (für Intervalle),

Union-Split-Find: Split(x) markiere  $x$ , Union(x) entfernt Markierung, Find(x) nächste Markierung rechts von  $x$

## 1.2 Wörterbücher

Wir kennen balancierte Suchbäume.

Problem: Teilmenge  $S \subseteq U$  mit  $U$  Universum, eventuell linear geordnet.

Speichere  $S$  (Schlüssel) in einer Datenstruktur  $D$  mit:

D.insert( $x$ )  $x \in U$ ,  $S \leftarrow S \cup \{x\}$

D.delete( $x$ )  $x \in U$ ,  $S \leftarrow S \setminus \{x\}$

D.lookup( $x$ )  $x \in U$ , testet ob  $x \in S$

Es soll zusätzlich zum Schlüssel Zusatzinformation gespeichert werden.

Wir behandeln 2 Datenstrukturen: Randomisierte Suchbäume, Perfektes Hashing

### 1.2.1 Randomisierter Suchbaum

Entwickelt von Seidel/Aragon

Idee: Verwende einen Zufallsprozess zur Balancierung von binären Suchbäumen.

Vorteile: Sehr einfache Implementierung, geringer Aufwand zur Verwaltung der Balance, effizient

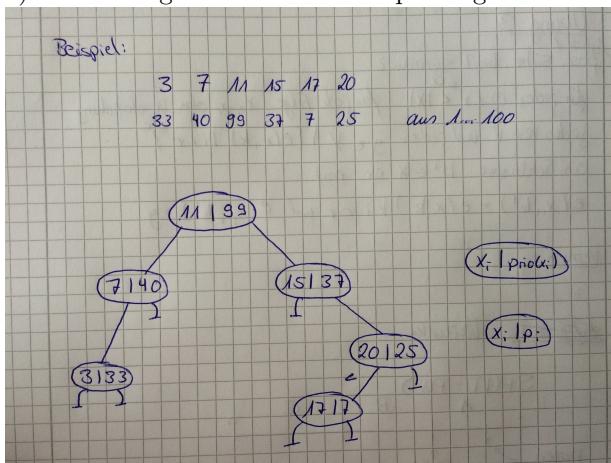
**Definition** RST(Randomized Search Tree):

Sei  $S = \{x_1, \dots, x_n\}$  eine Menge von  $n$  Schlüsseln aus einem linear geordneten Universum  $U$ .

Jedem  $x_i \in S$  wird zusätzlich eine Zufallszahl  $prio(x_i)$ , seine Priorität zugeordnet. Diese Zufallszahlen sind gleichverteilte reelle Zahlen aus  $[0,1]$  oder praktisch ganze Zahlen aus  $[0, \dots, 2^{32} - 1]$ .

Ein RST für  $S$  ist:

- 1) Ein Knotenorientierter binärer Suchbaum für die Paare  $(x_i, prio(x_i))$
- 2) Gleichzeitig ein Maximum-Heap bezüglich der Prioritäten  $prio(x_i)$



Die Prioritäten auf einem Pfad von der Wurzel zu einem Blatt sind monoton fallend. Die maximale Priorität ist in der Wurzel.

- 1) Sie  $x_i$  der Schlüssel mit maximaler Priorität, Erzeuge den Wurzelknoten  $v$  mit dem Inhalt  $(x_i, p_i)$  mit  $p_i$  maximal.
- 2)  $v.left \leftarrow (\{x_j | x_j < x_i\})$
- 3)  $v.right \leftarrow (\{x_k | x_k > x_i\})$

Beobachtung: Degenerierter Baum ist unwahrscheinlich, da er nur auftritt, wenn in der Liste der maximale Schlüssel immer die maximale Priorität hat.

Erwartete Höhe (Kosten der Operationen) ist  $\mathcal{O}(\log(n))$ .

Alternative Konstruktion:

Füge alle Schlüssel in absteigender Reihenfolge bzgl. der Priorität in einen normalen Knoten-orientierten Baum ein. Dabei fügt das Insert den entsprechenden Knoten an die richtige Position bzgl. des Schlüssels ein. Der Baum wird durch eine zufällige (weil Prioritäten zufällig) Folge von Inserts aufgebaut.

Operationen auf einem RST:

- 1) Lookup( $x$ ): Normale Suche im binären Suchbaum. Falls  $x \in S$  gilt:  $\mathcal{O}(Tiefe(x))$ , also maximal  $\mathcal{O}(Hoehe(T))$
- 2) Insert( $x$ ): Bestimme eine zufällige Priorität  $prio(x) \in [0, 1]$ . Füge einen neuen Knoten (Blatt)  $v$  mit Inhalt  $(x, prio(x))$  gemäß dem Schlüssel  $x$  durch ein normales Insert in den RST ein. Im Allgemeinen wird hier die Heapeigenschaft bzgl. der Prioritäten verletzt. Rotiere deswegen  $v$  nach oben, bis  $prio(vater(v)) \geq prio(v)$  oder bis  $v = \text{Wurzel}$

Beim Rotieren gibt es zwei Fälle:

1.  $v$  ist rechtes Kind  $\rightarrow \text{rotate\_left}(vater(v))$
2.  $v$  ist linkes Kind  $\rightarrow \text{rotate\_right}(vater(v))$

Dabei sind es meistens nur wenige auszuführende Rotationen. Der Fall, dass bis zur Wurzel rotiert werden muss ( $prio(v)$  ist neues Maximum) ist selten.

- 3) Delete( $x$ ): Lookup( $x$ ) liefert Knoten  $v$  mit Schlüssel  $x$ . Rotiere  $v$  nach unten, bis er ein Blatt ist. Lösche diesen dann.

Runterschieben des zu löschenen Knotens:

```

//Sei pl die Prioritaet des linken Kindes , pr die des rechten .
while v ist kein Blatt do
    if pl > pr und v.right == null
        rotate_right(v)
    else
        rotate_left(v)
fi
od

```

4) Split(y) teilt den Unterbaum von y.  $s_1 = \{x \in S | x \leq y\}$ ,  $s_2 = \{x \in S | x > y\}$

5) Join( $T_1, T_2$ ): Bildet aus den beiden uebergebenen RST einen neuen. ( $T_1, T_2$  sind RST von  $s_1, s_2$ )

Bedingung:  $\max(s_1) < \min(s_2)$

Kontruiere RST mit  $\max(s_1) < x < \min(s_2)$  mit x als Wurzelement mit unendlicher Prioritaet. Entferne x mit delete(x).

Laufzeitanalyse:

Wir analysieren die erwarteten Kosten einer Delete-Operation in einem RST mit n Knoten. D.h. fuer Schlüssel  $x_1, \dots, x_n$ , die durch Insert eingefügt wurden. Beobachtung: Insert ist das inverse Delete

Sei T ein RST für die Menge  $S = \{x_1, \dots, x_n\}$  mit  $x_1 < \dots < x_n$  entstanden durch Folge von Inserts.

Betrachte Operation Delete( $x_k$ ) mit  $1 \leq k \leq n$

Allgemeine Situation:  $P_k$ : Suchpfad nach  $x_k$ ,  $L_k$ : Rechtes Rückgrat vom linken Unterbaum von  $x_k$ ,  $R_k$ : linkes Rückgrat vom rechten Unterbaum von  $x_k$ .

Kosten:  $\mathcal{O}(|P_k| + |L_k| + |R_k|)$  mit  $P_k$  Lookup,  $L_k$  und  $R_k$  das Rotieren.

Lemma 1: Sei  $S = \{x_1, \dots, x_n\}$  mit  $x_i < x_{i+1}$  für  $i = 1, \dots, n-1$  und  $\{\text{prio}(x_i) | i = 1, \dots, n\}$  eine Menge von gleichverteilten reelen Zufallszahlen aus  $[0,1]$  abgespeichert in einem RST (Betrachte den Knoten v, der einen beliebigen Schlüssel  $x_k$  enthält). Dann gilt:

a)  $E(|P_k|) = H_k + H_{n-k+1} - 1$  H: Harmonische Zahl.

b)  $E(|L_k|) = 1 - \frac{1}{k}$

c)  $E(|R_k|) = a - \frac{1}{n-k+1}$

Wobei  $H_k = \sum_{i=1}^k \frac{1}{i}$

1)  $H_k \leq 1 + \ln(k)$

2)  $\sum_{i=0}^{k-1} H_i = k * (H_k - 1)$

Beweis Lemma 1:

Betrachte eine Permutation  $\Pi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$ , die S nach Prioritäten absteigend sortiert, d.h.  $\text{prio}(\text{x}_{\Pi(1)}) > \text{prio}(\text{x}_{\Pi(2)}) \dots$ . Dann gilt (Beobachtung):

1) Jede der Permutationen  $\Pi$  ist gleich wahrscheinlich (da Prioritäten gleichverteilt).

2) Man erhält den RST (denselben binären Baum) durch normales Einfügen von  $x_{\Pi(1)}, \dots, x_{\Pi(n)}$  in einen normalen binären Suchbaum.

3) Dann wächst der Baum nur an den Blättern.

Durch die zufälligen Prioritäten ist der RST immer ein zufälliger Baum.

Teil a): Suchpfad  $P_k$

Seien  $P'_k$  und  $P''_k$  eine Zerlegung von  $P_k$ :  $P'_k = \{v \in P_k | \text{key}(v) \leq x_k\}$ ,  $P''_k = \{v \in P_k | \text{key}(v) \geq x_k\}$

$w \in P''_k$  war zum Zeitpunkt seiner Einfügung der kleinste aller Knoten mit Schlüssel  $\geq x_k$ ,  $v \in P_k$  der größte  $\leq x_k$

Wir zeigen  $E(|P'_k|) = H_k$ , sowie  $E(|P''_k|) = H_{n-k+1}$ , daraus folgt a).

Zur Abschätzung von  $E(|P'_k|)$  definieren wir ein Spiel:

Spiel A: Ziehe zufällige Schlüssel aus  $\{x_1, \dots, x_k\}$  und zähle wie oft der Schlüssel maximal ist. Das Spiel endet, sobald  $x_k$  gezogen wird.

$A^k$  = erwartetes Ergebnis von Spiel A für den Schlüssel  $x_k = E(|P + k'|)$

Induktion:  $A^i = H_i$  für  $i < k$  (Erwartungswert:  $E = \sum_X (\text{prob}(x) * \text{value}(x))$ )

$A^k = \frac{1}{k} * 1 + \sum_{i=1}^{k-1} \frac{1}{k} * (1 - A^{k-i}) = \frac{1}{k} \sum_{i=1}^k (1 + A^{k-i}) = \frac{1}{k} (k + \sum_{i=0}^{k-1} H_i) = 1 + \frac{1}{k} (k(H_k - 1)) = H_k$

$\frac{1}{k} * 1$ : Im ersten Zug  $x_k$ ,  $\sum_{i=1}^{k-1} \frac{1}{k} * (1 - A^{k-i})$  im 1. Zug nicht  $x_k$  sondern  $x_i$  aus  $\{x_1, \dots, x_{k-1}\}$

Spiel B: Kandidaten  $\{x_k, \dots, x_n\}$ . Ziehe zufällige Elemente und zähle wie oft ein neues Minimum gezogen wird.

$B^k$  = Erwartungswert dieser Zahl

Behauptung:  $B^k = H_{n-k+1}$ , Beweis symmetrisch zu A (Übung)

$\Rightarrow$  Teil a) des Lemmas:  $E(|P_k|) = H_k + H_{n-k+1} - 1$ , weil  $x_k$  doppelt gezählt.

Wie in Teil a) können wir annehmen, dass der RST ein normaler binärer Baum mit zufälliger Insert-Reihenfolge

$x_{\Pi}(1), \dots, x_{\Pi}(n)$  ist.

$L_k$ : Ziehe zufällige Elemente sobald  $x_k$  gezogen ist (Trigger). Zähle wie oft ein Element  $> x_k$  gezogen wird, das maximal ist.

$R_k$ : symmetrisch mit Element  $< x_k$  minimal.

Spiel C: k Kandidaten  $\{x_1, \dots, x_k\}$

$$C^k := E(|L_k|) = \frac{1}{k} * A^{k-1} + \sum_{i=1}^{k-1} (\frac{1}{k} C^{k-i})$$

$$C^k = \frac{1}{k} (H_{k-1} + \sum_{i=0}^{k-1} C^i)$$

Trick: schätze  $\Delta_j := C^{j+1} - C^j$  ab  $\Rightarrow \sum_{i=0}^{k-1} C^i = \sum_{i=1}^{k-1} (\Delta_j)$

Betrachte:  $(j+1) * C^{j+1} - j * C^j = C^j + H_j - H_{j-1}$

$$\Rightarrow (j+1) * (C^{j+1} - C^j) = H_j - H_{j-1}$$

$$\Rightarrow \frac{1}{j*(j+1)} = C^{j+1} - C^j$$

$$\Rightarrow \Delta_j = \frac{1}{j} - \frac{1}{j+1}$$

$$C^k = \sum_{j=1}^{k-1} \Delta_j = \sum_{j=1}^{k-1} (\frac{1}{j} - \frac{1}{j+1}) = 1 - \frac{1}{k}$$

Spiel D: Kandidaten  $\{x_k, \dots, x_n\}$ , zähle wie oft ein Element nach  $x_k$  gezogen wird, das minimal ist.

$$D^k = \frac{1}{n-k+1} B^{k-1} + \sum_{i=k+1}^n (\frac{1}{n-k+1} D^{i-k})$$

Satz: Sei T ein RST für eine Menge von n Schlüsseln.

1. Die erwartete Laufzeit für Insert, Delete und Lookup ist jeweils  $\mathcal{O}(\log(n))$

2. Die erwartete Zahl der Rotationen bei Insert oder Delete ist  $< 2$

Beweis:

1. Kosten für Lookup:  $E(|P_k|)$  nach einem  $x_k$

$$\text{Lemma Teil a)} E(|P_k|) = H_k + H_{n-k+1} - 1$$

$$\Rightarrow \leq 2H_n = \mathcal{O}(\ln(n)) = \mathcal{O}(\log(n))$$

Insert, Delete von  $x_k$

$$\text{Kosten } \mathcal{O}(|P_k| + |R_k|) = \mathcal{O}(H_k + H_{n-k+1} - 1 + 1 - \frac{1}{k} + 1 - \frac{1}{n-k+1}) = \mathcal{O}(H_n) = \mathcal{O}(\log(n))$$

2. Erwartete Anzahl an Rotationen:

$$E(|L_k|) + E(|R_k|) = 1 - \frac{1}{k} + 1 - \frac{1}{n-k+1} < 2$$

## 1.3 Hashing

Speicherung dünn besetzter Tabellen / sparse tables

Spezielle Wörterbücher, da Schlüssel ganze Zahlen. Es werden keine Vergleiche ( $\leq, \geq$  etc.) auf der Schlüsselmenge durchgeführt. Es gibt keine lineare Ordnung.

Genauer: Verwaltet Schlüsselmenge  $S \subseteq \{0, \dots, N-1\}$  mit eventuell dazugehörigen Daten (Satellitendaten).

Operationen: Lookup(x), Insert(x), Delete(x)

Wie immer gilt  $n = |S|$ ,  $N = \text{Anzahl aller Schlüssel}$ .

Triviale Lösung:

Verwende ein Feld  $T[0 \dots N-1]$  (Tafel). Speichere  $x \in S$  (mit seinen Daten) an Tafelposition x, d.h.  $T[x] \leftarrow x$

Für alle  $y \notin S : T[y] \leftarrow \text{null}(-1)$

Lookup(x): return  $T[x]$

Problem: Speicherplatz  $\mathcal{O}(N)$ , Ziel  $\mathcal{O}(n)$

Ziel Hashing: Laufzeit  $\mathcal{O}(1)$ , Speicher  $\mathcal{O}(n)$

Hashtable:  $T[0 \dots m-1]$  mit m Größe der Tafel  $m \ll N$

Hashfunktion:  $h : U \rightarrow \{0, \dots, m-1\}$  mit Universum  $U = \{0, \dots, N-1\}$ . Speichert  $x \in S$  an Position  $h(x)$

Insert:  $T[h(x)] \leftarrow x$  (Daten)

Lookup(x): Teste ob  $T[h(x)] == x$

Die Hashfunktion entscheidet, wie die Tabelle kleiner als die Schlüsselmenge werden kann.

Häufig verwendete naive Hashfunktion:  $h : x \rightarrow x \bmod m$

Es treten Kollisionen auf, wenn  $h(x) = h(y)$  mit  $x \neq y$

Kollisionsbehandlung:

1. Hashing mit Verkettung (Kollisionsmengen werden in Liste gespeichert und mit abgefragtem Wert abgeglichen)  
Oft  $m < n$ . Speichere alle Schlüssel  $x \in S$  mit  $h(x) = i$  in einer Liste  $T[i]$ . Meist wird als Funktion der einfache Modulo verwendet.

Lookup(x): Durchsuche Liste von  $T[h(x)]$  linear.  $\mathcal{O}(1 + |T[h(x)]|)$ , worstcase  $\mathcal{O}(n)$

Erwartete Kosten:  $\mathcal{O}(1 + \frac{n}{m})$  (Übung), Belegungsfaktor  $\beta = \frac{n}{m}$  (Erwartete Länge einer Liste  $T[x]$ )

Insert(x): Falls Lookup(x)=null füge x an erste Stelle von  $T[h(x)]$  ein.

Delete(x): Entfernt x aus der Liste  $T[h(x)]$

Verbesserung: Immer wenn  $\beta > 4$  wird, verdopple die Tafelgröße. 1 sehr teures Insert  $\rightarrow$ , im Schnitt weiter  $\mathcal{O}(1)$

Bei Delete und kleinem  $\beta$  kann Tabelle halbiert werden. 1 sehr teures Delete.

2. Hashing mit offener Adressierung (Ausprobieren einer Folge von Positionen)

Voraussetzung:  $n \leq m$  und damit  $\beta \leq 1$

Idee: Folge von Hashfunktionen  $h_0, h_1, \dots: h_i(x) = (f(x) + i * g(x)) \bmod m$

Mit f,g Hashfunktionen  $U \rightarrow \{0, \dots, m-1\}$

f(x) gibt die Startposition an, g(x) verschiebt diese.

$h(x)=1$  heißt Linear Probing.

Falls belegt probiere  $h_1(x), h_2(x), \dots$ , bis freie Stelle gefunden.

Der Status der Positionen kann in einem zweiten Feld  $\text{status}[0..m-1]$  gespeichert werden (frei, besetzt, gelöscht).

Lookup(x): Durchsuche Folge  $T[h_0(x)], T[h_1(x)] \dots$  bis x gefunden, oder freie Position (dann ist x nicht enthalten)

Delete(x): Lookup(x) , markiere Position als frei. Problem: Elemente dahinter unerreichbar.

Lösung: gelöscht flag im Statusarray. Lookup übergeht diesen und hält nicht, Insert erkennt es als freies Feld.

3. Perfektes Hashing (keine Kollision durch injektive Funktion) Voraussetzung  $n \leq m$

$S \subseteq \{0, \dots, N-1\}$   $n = |S|$  verwende Tafel der Größe  $m \geq n$  und  $m = \mathcal{O}(n)$

Statisch: S ist fest, nur 2 Operationen. Init(S) Konstruktor, Lookup(x)

Ziel: Tafelgröße  $s = \mathcal{O}(n)$ , Hashfunktion injektiv auf S

Gegeben  $S \subseteq \{0, \dots, N-1\}$  und Tafel  $T[-, \dots, s-1]$

Injektive Funktion  $h: \{0, \dots, N-1\} \rightarrow \{0, \dots, s-1\}$  injektiv auf S

Idee: Verwende ein randomisiertes Verfahren, d.h. wähle eine zufällige Funktion aus den Kandidaten.

Originalarbeit: Storing a Sparse Table with  $\mathcal{O}(1)$  worst-case Access Time

Verwende zweistufiges Hashing Schema (injektiv), Auswahl der Funktion durch Randomisierung

1. Schritt: Funktion (muss noch nicht injektiv sein) bildet auf eine Liste von Buckets  $(W_0, \dots, W_{s-1})$   $W_i = \{x \in S \mid h(x) = i\}$  ab, mit s der Größe der ersten Stufe.

2. Schritt: Für jedes  $W_i$  gibt es eine eigene Hashfunktion  $h_i$ , die jeweils auf eine weitere Tafel der Größe s abbilden.  $h_i$  ist injektiv auf  $W_i$ . Für jedes  $W_i$  gibt es eine eigene 2. Tafel  $\{m_0, \dots, m_{s-1}\}$ . Dadurch gibt es auf der zweiten Stufe quadratische Tafelgröße.

Man findet "leicht" eine Funktion h der ersten Stufe mit  $\sum_{i=0}^{s-1} |W_i|^2 = \mathcal{O}(n)$

Für Tafeln quadratischer Größe findet man relativ leicht injektive Hashfunktion. Sei p eine Primzahl mit  $p > N$ .

$U = \{0, \dots, N-1\}, S \subseteq U, n = |S|, s \in \mathbb{N}$  Tafelgröße

Betrachte folgende Hashfunktionen  $h_1, \dots, h_{p-1}$

$h_k: \{0, \dots, N-1\} \rightarrow \{0, \dots, s-1\}$

$h_k(x) = (k * x \bmod p) \bmod s$

Jede Funktion  $h_k, 1 \leq k \leq p-1$  verteilt die Menge S auf s Buckets:

$W_0^k, \dots, W_{s-1}^k$ , d.h. genauer:  $W_i^k = \{x \in S \mid h_k(x) = i\}$

Lemma1: Für jede Menge  $S \subseteq \{0, \dots, N-1\}$  mit  $|S| = n$  gilt:

$\exists k, 1 \leq k \leq p-1$  mit  $\sum_{i=0}^{s-1} (\binom{|W_i^k|}{2}) < \frac{n^2}{s}$

$\binom{n}{k} = \frac{n!}{k!(n-k)!}$ : Anzahl der k-elementigen Teilmengen einer Menge der Größe n

$\binom{n}{2} = \frac{n \cdot n - 1}{2}$

hier:  $\binom{|W_i^k|}{2}$ : Anzahl aller  $\{x, y\} x \neq y, x, y \in S$ , die im selben Bucket  $W_i^k$  landen. (#Kollisionen) Beweis:

Behauptung:  $\sum_{k=1}^{p-1} \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} < (p-1) \frac{n^2}{s}$  (Beweis später)

Daraus folgt Lemma1. Indirekt gilt Lemma 1 nicht  $\Rightarrow \forall 1 \leq k \leq p-1 : \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} \geq \frac{n^2}{s} \Rightarrow \sum_{k=1}^{p-1} \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} \geq (p-1) \frac{n^2}{s}$  Widerspruch zu Behauptung

Folgerung 1: Für  $s = n$  (d.h. Tafelgröße n) folgt aus Lemma1:  $\exists k, 1 \leq k \leq p-1 : \sum_{i=0}^{s-1} (|W_i^k|^2) < 3n$

Beweis: Betrachte Lemma1 für  $s=n$

$\exists 1 \leq k \leq p-1 : \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} < n$

$\sum_{i=0}^{s-1} \frac{|W_i^k| * (|W_i^k| - 1)}{2} < n$

$\sum_{i=0}^{s-1} |W_i^k|(|W_i^k| - 1) < 2n$

$$\sum_{i=0}^{n-1} (|W_i^k|^2 - |W_i^k|) < 2n$$

$$\sum_{i=0}^{n-1} |W_i^k|^2 < 2n + \sum_{i=0}^{n-1} |W_i^k| \text{ wobei } \sum_{i=0}^{n-1} |W_i^k| = n \text{ also } |S|$$

Folgerung 2: Für  $s = n^2$  folgt aus Lemma 1:  $\exists 1 \leq k' \leq p-1$  sodass die Hashfunktion  $h_{k'} : x \rightarrow (k' * x \bmod p) \bmod n^2$  die injektiv auf  $S$  ist, d.h.  $|W_i^{k'}| \leq 1$  für  $i = 0, \dots, n^2-1$

Für quadratische Tafelgrößen existiert eine perfekte Hashfunktion  $h_{k'}$

Beweis: Betrachte Lemma 1 für  $s = n^2$

$$\exists 1 \leq k' \leq p-1 \text{ mit } \sum_{i=0}^{n-1} \binom{|W_i^{k'}|}{2} < \frac{n^2}{2} = 1 \Rightarrow \sum_{i=0}^{n^2-1} \binom{|W_i^{k'}|}{2} = 0 \Rightarrow \forall 0 \leq i \leq n^2-1 : |W_i^{k'}| \leq 1 \Rightarrow \#Kollisionen = 0$$

Vermeidung des quadratischen Speicherplatzes durch ein 2-Stufiges hashing-Schema.

$$1. \text{ Stufe: Verwende Tafelgröße } s=n \text{ und wähle ein } k \text{ gemäß der Folgerung d.h. } \sum_{i=0}^{n-1} |W_i^k|^2 < 3n$$

Die Hashfunktion  $h_k(x) : x \rightarrow (kx \bmod p) \bmod n$  verteilt  $s$  auf eine Tafel der Größe  $n$ , sodass die Summe der Quadrate der Bucketgrößen kleiner ist als  $3n$ .

2. Stufe: Für jedes nicht-leere Bucket  $W_i^k$  der 1. Stufe verwende eine Tafel der Größe  $s_i = |W_i^k|^2$  und wähle  $k_i$  gemäß Folgerung 2. Genauer: Für  $i=0, \dots, n-1$  wähle ein  $k_i$  sodass  $h_{k_i}(x) : x \rightarrow (k_i x \bmod p) \bmod s_i$  injektiv auf  $W_i^k$  ist.

Gesamtbedarf: 1. Stufe: Platz  $n$ , 2. Stufe:  $\sum_{i=0}^{n-1} |W_i^k|^2 < 3n$  ungefähr  $4n$ , genauer später.

Problem: Wie findet man diese injektiven Funktionen, d.h. wie wählt man  $k$  und  $k'$

Idee: Erhöhe die Tafelgrößen auf Stufe 1 und Stufe 2 um einen konstanten Faktor. Dann erfüllen mindestens 50% aller  $k$  die Bedingungen von Folge 1&2

Beweis der Behauptung:

$$(*) \sum_{k=1}^{p-1} \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} < (p-1) \frac{n^2}{s} \Rightarrow \exists k : \sum_{i=0}^{s-1} \binom{|W_i^k|}{2} < \frac{n^2}{s}$$

Mit  $p$  Primzahl  $p \geq N$  Die Summe ist: Anzahl aller Paare  $(l, \{x, y\})$  mit  $1 \leq l \leq p-1$ ,  $x, y \in S$ ,  $x \neq y$  und  $h_k(x) = h_k(y)$  (Kollision).

Beitrag eines festen Paares  $x \neq y$  zu der Summe = Anzahl aller  $k$  mit  $h_k(x) = h_k(y)$

$$(kx \bmod p - ky \bmod p) \bmod s = 0$$

$$kx \bmod p - ky \bmod p = i * s \quad i \in \mathbb{Z}$$

$k * (x - y) \bmod p \in \{-(p-1), \dots, p-1\}$  Vielfaches von  $s$  aus der Menge  $\Rightarrow \frac{2*(p-1)}{s}$  verschiedene Gleichungen  
Lösen der Gleichung nach  $k$   $k = \frac{is}{x-y} \bmod p$  Da  $p$  eine Primzahl ( $\mathbb{Z}_p$  ein Körper) existiert eine wohl-definierte Division (inverse zur Multiplikation) und daher besitzt jede Gleichung höchstens eine Lösung für  $k$ .

$$\Rightarrow \text{Der Beitrag von } \{x, y\} x \neq y \text{ zur Summe } \leq \frac{2(p-1)}{s}$$

Wir summieren über alle 2-elementigen Teilmengen  $\{x, y\}, x \neq y, x, y \in S$ :

$$\text{Anzahl der Teilmengen } (*) \leq \binom{n}{2} * \frac{2(p-1)}{s} = \frac{n(n-1)}{2} * \frac{2(p-1)}{s} \leq n^2 \frac{p-1}{s}$$

Implementierung Beispiel:

3 Felder (1. Stufe)  $W[0..n-1]$  mit  $W[i]$  Pointer auf Bucket 2. Stufe.

$\text{size}[0..n-1]$  Anzahl aller Elemente in  $W_i^k$

$K[0..n-1]$  mit  $K[i] = k_i$   $k$ -Werte der 2. Stufe.

Variable  $k_0 = k$  der ersten Stufe

Für 2. Stufe:  $n$  Tafeln  $B_i$  mit  $i = 0..n-1$  und  $B_i[0..size[i]^2 - 1]$

Platzbedarf:  $\text{size} + K + W + k_0 + B_i = 3n + 1 + \sum_{i=0}^{n-1} \text{size}[i]^2 \leq 6n + 1 = \mathcal{O}(n)$

Speichere  $x \in S$  wie folgt:  $i \leftarrow h_{k_0}(x)$ ,  $j \leftarrow h_{K[i]}(x)$ ,  $W[i][j] \leftarrow x$

Mit  $h_{k_0} = ((k_0 * x \bmod p) \bmod n)$  und  $h_{K[i]}(x) = ((K[i] * x) \bmod p) \bmod size[i]^2$

Lookup: Teste ob  $W[i][j] == x$

Bemerkung: Für  $\text{size}[i]=0$  kein  $B_i$  auf 2. Stufe,  $W[i]=\text{null}$ ; Für  $\text{size}[i]=1$  keine neue Hashfunktion  $K[i]$ , speichere direkt.

Für kleine  $\text{size}[i]$  konstant kann man Verkettung anstelle von Hashfunktion+Tabelle nehmen.