15.1 Union Find

15.2 Kruskal’s algorithm – a quick reminder

15.2.0.1 Compute minimum spanning tree

(A) $G$: Undirected graph with weights on edges.
(B) $Q$: Compute MST (minimum spanning tree) of $G$.
(C) Kruskal’s Algorithm:
   (A) Sort edges by increasing weight.
   (B) Start with a copy of $G$ with no edges.
   (C) Add edges by increasing weight, and insert into graph $\iff$ do not form a cycle.
   (i.e., connect two different things together.)

15.2.0.2 Kruskal’s Algorithm

Process edges in the order of their costs (starting from the least) and add edges to $T$ as long as they don’t form a cycle.
MST of $G$: 
15.2.1 Requirements from the data-structure
15.2.1.1 Requirements from the data-structure

(A) Maintain a collection of sets.
(B) \texttt{makeSet}(x) - creates a set that contains the single element \(x\).
(C) \texttt{find}(x) - returns the set that contains \(x\).
(D) \texttt{union}(A, B) - returns set = union of \(A\) and \(B\). That is \(A \cup B\).
... merges the two sets \(A\) and \(B\) and return the merged set.

15.2.2 Amortized analysis
15.2.2.1 Amortized Analysis

(A) Use data-structure as a black-box inside algorithm.
... Union-Find in Kruskal algorithm for computing MST.
(B) Bounded worst case time per operation.
(C) Care: overall running time spent in data-structure.
(D) \textit{amortized running-time} of operation
   = average time to perform an operation on data-structure.
(E) Amortized time per operation = \(\frac{\text{overall running time}}{\text{number of operations}}\).

15.2.3 The data-structure

15.2.4 Reversed Trees

15.2.4.1 Representing sets in the Union-Find DS
The Union-Find representation of the sets $A = \{a, b, c, d, e\}$ and $B = \{f, g, h, i, j, k\}$. The set $A$ is uniquely identified by a pointer to the root of $A$, which is the node containing $a$.

15.2.5 Reversed Trees

15.2.5.1  Reversed Trees:

(A) Initially: Every element is its own node.
(B) Node $v$: $p(v)$ pointer to its parent.
(C) Set uniquely identified by root node/element.

(B) makeSet: Create a singleton pointing to itself:

(C) find($x$):

(A) Start from node containing $x$, traverse up tree, till arriving to root.
(B) $x \rightarrow b \rightarrow a$
(C) $a$: returned as set.

15.2.6 Union operation in reversed trees

15.2.6.1 Just hang them on each other.

union($a, p$): Merge two sets.

(A) Hanging the root of one tree, on the root of the other.
(B) A destructive operation, and the two original sets no longer exist.

15.2.6.2 Pseudo-code of naive version...

```
makeSet(x)
\[ p(x) \leftarrow x \]

find(x)
if $x = p(x)$ then
  return $x$
else
  return find($p(x)$)

union(x, y)
\[ A \leftarrow \text{find}(x) \]
\[ B \leftarrow \text{find}(y) \]
\[ p(B) \leftarrow A \]
```
15.2.7 Example...

15.2.7.1 The long chain

After:  \texttt{makeSet}(a), \texttt{makeSet}(b), \texttt{makeSet}(c), \texttt{makeSet}(d), \texttt{makeSet}(e), \texttt{makeSet}(f), \texttt{makeSet}(g), \texttt{makeSet}(h)

\texttt{union}(g, h)
\texttt{union}(f, g)
\texttt{union}(e, f)
\texttt{union}(d, e)
\texttt{union}(c, d)
\texttt{union}(b, c)
\texttt{union}(a, b)

15.2.7.2 Find is slow, hack it!

(A) \texttt{find} might require \(\Omega(n)\) time.
(B) Q: How improve performance?
(C) Two “hacks”:

(i) \textit{Union by rank}:

\begin{itemize}
  \item Maintain in root of tree, a bound on its depth (\textit{rank}).
  \item \textbf{Rule}: Hang the smaller tree on the larger tree in \texttt{union}.
\end{itemize}

(ii) \textit{Path compression}:

\begin{itemize}
  \item During find, make all pointers on path point to root.
\end{itemize}

15.2.7.3 Path compression in action...

(a) The tree before performing \texttt{find}(z), and (b) The reversed tree after performing \texttt{find}(z) that uses path compression.
15.2.7.4 Pseudo-code of improved version...

\begin{align*}
\text{makeSet}(x) & \quad \text{\begin{tabular}{l}
\hspace{1cm} p(x) \leftarrow x \\
\hspace{1cm} \text{rank}(x) \leftarrow 0
\end{tabular}} \\
\text{find}(x) & \quad \text{\begin{tabular}{l}
\hspace{1cm} \text{if } x \neq \overline{p}(x) \text{ then} \\
\hspace{2cm} \overline{p}(x) \leftarrow \text{find}(\overline{p}(x)) \\
\hspace{1cm} \text{return } \overline{p}(x)
\end{tabular}} \\
\text{union}(x, y) & \quad \text{\begin{tabular}{l}
\hspace{1cm} A \leftarrow \text{find}(x) \\
\hspace{1cm} B \leftarrow \text{find}(y) \\
\hspace{1cm} \text{if } \text{rank}(A) > \text{rank}(B) \text{ then} \\
\hspace{2cm} \overline{p}(B) \leftarrow A \\
\hspace{1cm} \text{else} \\
\hspace{2cm} \overline{p}(A) \leftarrow B \\
\hspace{1cm} \text{if } \text{rank}(A) = \text{rank}(B) \text{ then} \\
\hspace{2cm} \text{rank}(B) \leftarrow \text{rank}(B) + 1
\end{tabular}}
\end{align*}

15.3 Analyzing the Union-Find Data-Structure

15.3.0.1 Definition

Definition 15.3.1. \( v \):: Node UnionFind data-structure \( D \)
\( v \) is leader \( \iff \) \( v \) root of a (reversed) tree in \( D \).

“When you’re not leader, you’re little people.”
“You know the score pal. If you’re not cop, you’re little people.” - Blade Runner (movie).

15.3.0.2 Lemma

Lemma 15.3.2. Once node \( v \) stop being a leader, can never become leader again.

Proof: (A) \( x \) stopped being leader because union operation hanged \( x \) on \( y \).
(B) From this point on...
(C) \( x \) might change only its parent pointer (find).
(D) \( x \) parent pointer will never become equal to \( x \) again.
(E) \( x \) never a leader again.

15.3.0.3 Another Lemma

Lemma 15.3.3. Once a node stop being a leader then its rank is fixed.

Proof: (A) rank of element changes only by union operation.
(B) union operation changes rank only for...
\( \text{the “new” leader of the new set.} \)
(C) if an element is no longer a leader, than its rank is fixed.

15.3.0.4 Ranks are strictly monotonically increasing

Lemma 15.3.4. Ranks are monotonically increasing in the reversed trees...
\( \hspace{1cm} \text{...along a path from node to root of the tree.} \)
15.3.0.5 Proof...

(A) Claim: $\forall u \rightarrow v$ in DS: $\text{rank}(u) < \text{rank}(v)$.
(B) Proof by induction. Base: all singletons. Holds.
(C) Assume claim holds at time $t$, before an operation.
(D) If operation is $\text{union}(A, B)$, and assume that we hanged $\text{root}(A)$ on $\text{root}(B)$.
   Must be that $\text{rank}(\text{root}(B))$ is now larger than $\text{rank}(\text{root}(A))$ (verify!).
   Claim true after operation!
(E) If operation $\text{find}$: traverse path $\pi$, then all the nodes of $\pi$ are made to point to the last node $v$ of $\pi$.
   By induction, $\text{rank}(v) > \text{rank}$ of all other nodes of $\pi$.
   All the nodes that get compressed, the rank of their new parent, is larger than their own rank. ■

15.3.0.6 Trees grow exponentially in size with rank

Lemma 15.3.5. When node gets rank $k \implies$ at least $\geq 2^k$ elements in its subtree.

Proof: (A) Proof is by induction.
(B) For $k = 0$: obvious since a singleton has a rank zero, and a single element in the set.
(C) node $u$ gets rank $k$ only if the merged two roots $u, v$ has rank $k - 1$.
(D) By induction, $u$ and $v$ have $\geq 2^{k-1}$ nodes before merge.
(E) merged tree has $\geq 2^{k-1} + 2^{k-1} = 2^k$ nodes.

15.3.0.7 Having higher rank is rare

Lemma 15.3.6. # nodes that get assigned rank $k$ throughout execution of Union-Find DS is at most $n/2^k$.

Proof: (A) By induction. For $k = 0$ it is obvious.
(B) when $v$ become of rank $k$. Charge to roots merged: $u$ and $v$.
(C) Before union: $u$ and $v$ of rank $k - 1$
(D) After merge: $\text{rank}(v) = k$ and $\text{rank}(u) = k - 1$.
(E) $u$ no longer leader. Its rank is now fixed.
(F) $u, v$ leave rank $k - 1 \implies v$ enters rank $k$.
(G) By induction: at most $n/2^{k-1}$ nodes of rank $k - 1$ created.
   $\implies$ # nodes rank $k$: $\leq (n/2^{k-1})/2 = n/2^k$.

15.3.0.8 Find takes logarithmic time

Lemma 15.3.7. The time to perform a single $\text{find}$ operation when we perform union by rank and path compression is $O(\log n)$ time.

Proof: (A) rank of leader $v$ of reversed tree $T$, bounds depth of $T$.
(B) By previous lemma: max rank $\leq \lg n$.
(C) Depth of tree is $O(\log n)$.
(D) Time to perform $\text{find}$ bounded by depth of tree.
15.3.0.9 **log\(^*\) in detail**

(A) \(\log^*(n)\): number of times to take \(\log\) of number to get number smaller than two.

(B) \(\log^* 2 = 1\)

(C) \(\log^* 2^2 = 2\).

(D) \(\log^* 2^{2^2} = 1 + \log^*(2^2) = 2 + \log^* 2 = 3\).

(E) \(\log^* 2^{2^{2^2}} = \log^*(65536) = 4\).

(F) \(\log^* 2^{2^{2^{2^2}}} = \log^* 2^{65536} = 5\).

(G) \(\log^*\) is a monotone increasing function.

(H) \(\beta = 2^{2^{2^{2^{2^2}}}}\): huge number

For practical purposes, \(\log^*\) returns value \(\leq 5\).

15.3.0.10 **Can do much better!**

**Theorem 15.3.8.** For a sequence of \(m\) operations over \(n\) elements, the overall running time of the UnionFind data-structure is \(O((n + m) \log^* n)\).

(A) Intuitively: UnionFind data-structure takes constant time per operation...
   (unless \(n\) is larger than \(\beta\) which is unlikely).

(B) Not quite correct if \(n\) sufficiently large...

15.3.0.11 **The tower function...**

**Definition 15.3.9.** \(\text{Tower}(b) = 2^{\text{Tower}(b-1)}\) and \(\text{Tower}(0) = 1\).

- \(\text{Tower}(i)\): a tower of \(2^{2^{2^{\cdots^{2^2}}}}\) of height \(i\).

  Observe that \(\log^*(\text{Tower}(i)) = i\).

**Definition 15.3.10.** For \(i \geq 0\), let \(\text{Block}(i) = [\text{Tower}(i-1) + 1, \text{Tower}(i)]\); that is

\[
\text{Block}(i) = [z, 2^z - 1]
\]

for \(z = \text{Tower}(i-1) + 1\).

Also \(\text{Block}(0) = [0, 1]\). As such,

- \(\text{Block}(0) = [0, 1]\), \(\text{Block}(1) = [2, 2]\), \(\text{Block}(2) = [3, 4]\), \(\text{Block}(3) = [5, 16]\), \(\text{Block}(4) = [17, 65536]\), \(\text{Block}(5) = [65537, 2^{65536}]\)...

15.3.0.12 **Running time of find...**

(A) RT of \(\text{find}(x)\) proportional to length of the path from \(x\) to the root of its tree.

(B) ...start from \(x\) and we visit the sequence:

\[x_1 = x, x_2 = \text{p}(x_1), x_3 = \text{p}(x_2), \ldots, x_i = \text{p}(x_{i-1}), \ldots, x_m = \text{p}(x_{m-1}) = \text{root of tree}.\]

(C) \(\text{rank}(x_1) < \text{rank}(x_2) < \text{rank}(x_3) < \ldots < \text{rank}(x_m)\).

(D) RT of \(\text{find}(x)\) is \(O(m)\).

**Definition 15.3.11.** A node \(x\) is **in the \(i\)th block** if \(\text{rank}(x) \in \text{Block}(i)\).

(E) Looking for ways to pay for the \(\text{find}\) operation.

(F) Since other two operations take constant time...
15.3.0.13 Blocks and jumping pointers

(A) maximum rank of node \( v \) is \( O(\log n) \).
(B) # of blocks is \( O(\log^* n) \), as \( O(\log n) \in \text{Block}(c \log^* n) \), \( (c: \text{constant, say 2}) \).
(C) \textbf{find} \( (x) \): \( \pi \) path used.
(D) partition \( \pi \) into each by rank.
(E) Price of \textbf{find} length \( \pi \).
(F) node \( x \): \( \nu = \text{index}_B(x) \) index block containing \( \text{rank}(x) \).
(G) \( \text{rank}(x) \in \text{Block}(\text{index}_B(x)) \).
(H) \( \text{index}_B(x): \text{block of } x \)

15.3.0.14 The path of find operation, and its pointers

15.3.0.15 The pointers between blocks...

(A) During a \textbf{find} operation...
(B) \( \pi \): path traversed.
(C) Ranks of the nodes visited in \( \pi \) monotone increasing.
(D) Once leave block \( i \)th, never go back!
(E) charge visit to nodes in \( \pi \) next to element in a different block...
(F) to total number of blocks \( \leq O(\log^* n) \).

15.3.0.16 Jumping pointers

Definition 15.3.12. \( \pi \): path traversed by \textbf{find}.

(A) If for \( x \in \pi \), the node \( \bar{p}(x) \) is in a different block than \( x \), then \( x \rightarrow \bar{p}(x) \) is a \textit{jump between blocks}.
(B) jump inside a block is an \textit{internal jump} (i.e., \( x \) and \( \bar{p}(x) \) are in same block).
15.3.0.17 Not too many jumps between blocks

Lemma 15.3.13. During a single find(x) operation, the number of jumps between blocks along the search path is \( O(\log^* n) \).

Proof: (A) \( \pi = x_1, \ldots, x_m \): path followed by find.
(B) \( x_i = p(x_{i-1}) \), for all \( i \).
(C) \( 0 \leq \text{index}_B(x_1) \leq \text{index}_B(x_2) \leq \ldots \leq \text{index}_B(x_m) \).
(D) \( \text{index}_B(x_m) = O(\log^* n) \).
(E) Number of elements in \( \pi \) such that \( \text{index}_B(x) \neq \text{index}_B(p(x)) \).
(F) \ldots at most \( O(\log^* n) \).

15.3.0.18 Benefits of an internal jump

(A) \( x \) and \( p(x) \) are in same block.
(B) \( \text{index}_B(x) = \text{index}_B(p(x)) \).
(C) find passes through \( x \).
(D) \( r_{\text{bef}} = \text{rank}(p(x)) \) before find operation.
(E) \( r_{\text{aft}} = \text{rank}(p(x)) \) after find operation.
(F) By path compression: \( r_{\text{aft}} > r_{\text{bef}} \).
(G) \( \implies \) parent pointer \( x \) jumped forward...
(H) \ldots and new parent has higher rank.
(I) Charge internal block jumps to this “progress”.

15.3.1 Changing parents...

15.3.1.1 Your parent can be promoted only a few times before leaving block

Lemma 15.3.14. At most \( |\text{Block}(i)| \leq \text{Tower}(i) \) find operations can pass through an element \( x \), which is in the \( i \)th block (i.e., \( \text{index}_B(x) = i \)) before \( p(x) \) is no longer in the \( i \)th block. That is \( \text{index}_B(p(x)) > i \).

Proof: (A) parent of \( x \) incr rank every-time internal jump goes through \( x \).
(B) At most \( |\text{Block}(i)| \) different values in the \( i \)th block.
(C) \( \text{Block}(i) = [\text{Tower}(i-1) + 1, \text{Tower}(i)] \)
(D) Claim follows, as: \( |\text{Block}(i)| \leq \text{Tower}(i) \).

15.3.1.2 Few elements are in the bigger blocks

Lemma 15.3.15. At most \( n/\text{Tower}(i) \) nodes are assigned ranks in the \( i \)th block throughout the algorithm execution.

Proof: By lemma, the number of elements with rank in the \( i \)th block

\[
\leq \sum_{k \in \text{Block}(i)} \frac{n}{2^k} = \sum_{k=\text{Tower}(i-1)+1}^{\text{Tower}(i)} \frac{n}{2^k} \\
= n \cdot \sum_{k=\text{Tower}(i-1)+1}^{\text{Tower}(i)} \frac{1}{2^k} \leq \frac{n}{2^{\text{Tower}(i-1)}} = \frac{n}{\text{Tower}(i)}.
\]
15.3.1.3 Total number of internal jumps is $O(n)$

**Lemma 15.3.16.** The number of internal jumps performed, inside the $i$th block, during the lifetime of the union-find data-structure is $O(n)$.

*Proof:* (A) $x$ in $i$th block, have at most $|\text{Block}(i)|$ internal jumps...
(B) ... after that all jumps through $x$ are between blocks, by lemma...
(C) $\leq n/\text{Tower}(i)$ elements assigned ranks in the $i$th block, throughout algorithm execution.
(D) total number of internal jumps is $|\text{Block}(i)| \cdot \frac{n}{\text{Tower}(i)} \leq \text{Tower}(i) \cdot \frac{n}{\text{Tower}(i)} = n$.

15.3.1.4 Total number of internal jumps

**Lemma 15.3.17.** The number of internal jumps performed by the Union-Find data-structure overall is $O(n \log^* n)$.

*Proof:* (A) Every internal jump associated with block it is in.
(B) Every block contributes $O(n)$ internal jumps throughout time.
(By previous lemma.)
(C) There are $O(\log^* n)$ blocks.
(D) There are at most $O(n \log^* n)$ internal jumps.

15.3.1.5 Result...

**Lemma 15.3.18.** The overall time spent on $m$ find operations, throughout the lifetime of a union-find data-structure defined over $n$ elements, is $O((m + n) \log^* n)$.

**Theorem 15.3.19.** If we perform a sequence of $m$ operations over $n$ elements, the overall running time of the Union-Find data-structure is $O((n + m) \log^* n)$.

15.3.1.6 More on strange functions...

Idea: Define a sequence of functions $f_i(x) = f_{i-1}^{(x)}(0)$

<table>
<thead>
<tr>
<th>Function</th>
<th>Inverse function</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_1(x) = x + 2$</td>
<td>$g_1(y) = y - 2$</td>
</tr>
<tr>
<td>$f_2(x) = 2x$</td>
<td>$g_2(y) = y/2$</td>
</tr>
<tr>
<td>$f_3(x) = 2^x$</td>
<td>$g_3(y) = \log y$</td>
</tr>
<tr>
<td>$f_4(x) = \text{Tower}(x)$</td>
<td>$g_4(x) = \log^* x$</td>
</tr>
<tr>
<td>$f_5(x) = \ldots$</td>
<td>$g_5(x) = \ldots$</td>
</tr>
</tbody>
</table>

$f_2(x) = f_1(f_2(x - 1)) = 2x$ $f_3(x) = f_2(f_3(x - 1)) = 2^x f_4(x) = f_3(f_4(x - 1)) = \text{Tower}(x)$

$g_i(x) = \#$ of times one has to apply $g_{i-1}(-)$ to $x$ before we get number smaller than 2.

$A(n) = f_n(n)$: **Ackerman function.**

**Inverse Ackerman function:**

$\alpha(n) = A^{-1}(n) = \min i$ s.t. $g_i(n) \leq i$.

15.3.1.7 Union-Find: Tarjan result

**Theorem 15.3.20** ([Tarjan [1975]]). If we perform a sequence of $m$ operations over $n$ elements, the overall running time of the Union-Find data-structure is $O((n + m)\alpha(n))$.

(The above is not quite correct, but close enough.)
Bibliography