

The background of the cover is a solid blue color. It features a complex geometric pattern of concentric hexagons and parallel lines that create a sense of depth and movement, resembling a stylized circuit board or a modern architectural design. The pattern is more prominent on the left side and fades slightly towards the right.

CSC263

Data Structures and Analysis

SINAN LI

2023

CONTENTS

I Data Structure

7

1 | Chapter 1 Priority Queues and Heaps

- 1.1 Implementation 9
 - 1.1.1 Attempts 9
 - 1.1.2 Implementation 10
- 1.2 Operations 11
 - 1.2.1 INSERT 11
 - 1.2.2 FIND-MAX 11
 - 1.2.3 EXTRACT-MAX 12
 - 1.2.4 BUILD-MAX-HEAP 13

2 | Chapter 2 Dictionaries

- 2.1 Binary Search Trees 16
 - 2.1.1 INSERT 16
 - 2.1.2 SEARCH 17
 - 2.1.3 DELETE 17
- 2.2 Balanced Search trees 18
 - 2.2.1 INSERT 20
 - 2.2.2 DELETE 21
 - 2.2.3 Rebalancing 22
- 2.3 Hashing 22
 - 2.3.1 Direct Access Tables 23
 - 2.3.2 Hash Tables 23

3 | Chapter 3

Dynamic Array

4 | Chapter 4

Graphs

- 4.1 Graphs 27
 - 4.1.1 Graphs 27
 - 4.1.2 Breadth-First Search 28
 - 4.1.3 Depth-First Search 32
 - 4.1.4 Strongly connected 34
- 4.2 Minimum Spanning Trees 35
 - 4.2.1 Prim's Algorithm 35
 - 4.2.2 Kruskal's algorithm 37
- 4.3 Disjoint Sets 38
 - 4.3.1 Implementation 38

II Algorithms

43

5 | Chapter 5

Sorting

- 5.1 Heap Sort 45
- 5.2 Quick Sort 45
 - 5.2.1 Deterministic Quick Sort 45
 - 5.2.2 Randomized Quick Sort 46
- 5.3 Topology Sort 46
 - 5.3.1 Directed Acyclic Graphs 46
 - 5.3.2 Topological Sort 47

III Analysis

49

6 | Chapter 6

Average Case Analysis

7

Chapter 7

Amortized Analysis

- 7.1 Aggregated Method *56*
- 7.2 Accounting Method *57*

8

Chapter 8

Lower Bounds

- 8.1 Comparison-Based Algorithms *59*
- 8.2 Information Theory Lower Bounds *61*
- 8.3 Adversary / Adversarial Arguments *61*

IV Appendices

Part I

Data Structure

PRIORITY QUEUES AND HEAPS

1

Data

- Collection of elements
- Each element x has a priority
 $x.priority$

Operations

- $INSERT(Q, x)$
Add x to Q
Note: $x.priority$ can be non-unique
- $MAX(Q)$
Return the element with max priority
Note: Q is unchanged
- $EXTRACT-MAX(Q)$
Remove and return the element with
the max priority

1.1 Implementation

1.1.1 Attempts

Implementation 1: Unsorted Array / Linked List

- $INSERT$ takes $\Theta(1)$ time in the worst case
- MAX takes $\Theta(n)$ time in the worst case
- $EXTRACTMAX$ takes $\Theta(n)$ time in the worst case

Implementation 2: Sorted Array / Linked List

- INSERT takes $\Theta(n)$ time in the worst case
- MAX takes $\Theta(1)$ time in the worst case
- EXTRACTMAX takes $\Theta(1)$ time in the worst case

1.1.2 Implementation

We want to combine the advantages of both data structures by having a “partially sorted” ADT – a (binary) *heap*.

There are two kinds of binary heaps: max-heaps and min-heaps. In both kinds, the values in the nodes satisfy a **heap property**, the specifics of which depend on the kind of heap.

- Max heap property: the key of every node x is *larger* than or equal to the keys of its children. The largest element in a max-heap is stored at the root.
- Min heap property: the key of every node x is *smaller* than or equal to the keys of its children. The smallest element in a min-heap is stored at the root,

These are called *heap orders*. There is no ordering between the siblings. A max/min heap is valid if it is a nearly complete binary tree and it satisfies the max/min heap property.

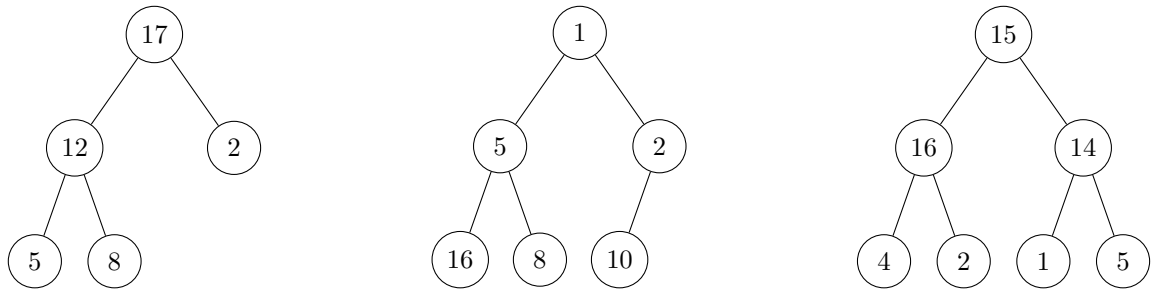


Figure 1.1: A valid max-heap (left), a valid min-heap (middle), and an invalid heap (right)

Although a heap is an **almost complete binary tree**¹, in practice, we usually use an array to store the data in memory. An array H that represents a heap is an object with two attributes: $H.length$, which (as usual) gives the number of elements in the array, and $H.heap-size$, which represents how many elements in the heap are stored within array H . The root of the tree is $H[1]$, and given the index i of a node, we can compute the indices of its parent, left child, and right child:

- | | | |
|-------------------------------------|--------------------|------------------------|
| • PARENT | • LEFT | • RIGHT |
| return $\lfloor i/2 \rfloor$ | return $2i$ | return $2i + 1$ |

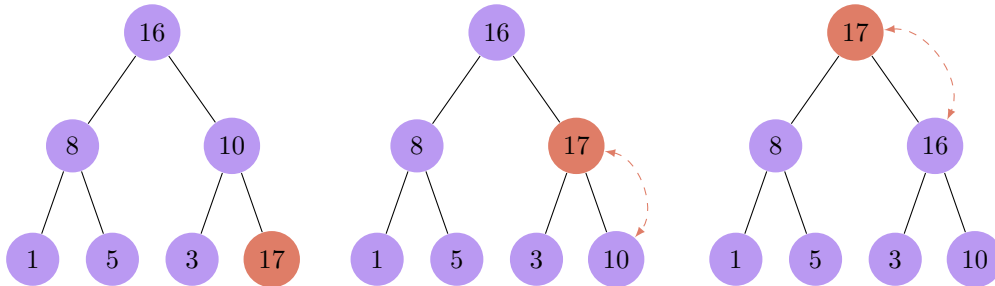
¹That is, the tree is completely filled on all levels except possibly the lowest, which is filled from the left up to a point.

1.2.1 INSERT

To insert element with key p into the heap H ,

- Increment $H.\text{heap-size}$ and add a new node with key p to the next available position
- Repeatedly swap the new item with its parent until the heap property is satisfied
This swapping process is called **bubbling up**
- Worst-case runtime: $\Theta(\lg n)$

For example, consider $\text{INSERT}(H, 17)$ where $H = [16, 8, 10, 1, 5, 3]$



```

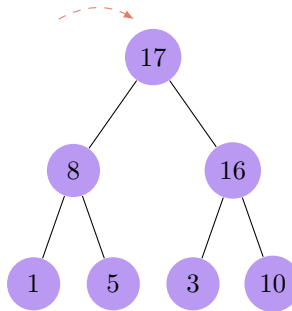
procedure MAX-HEAP-INSERT( $H, p$ )
   $i \leftarrow H.\text{heap-size} \leftarrow H.\text{heap-size} + 1$ 
   $H[i] = p$ 
  while PARENT( $i$ ) > 0 and  $H[i] > H[\text{PARENT}(i)]$  do
    swap  $H[i]$  with  $H[\text{PARENT}(i)]$ 
     $i \leftarrow \text{PARENT}(i)$ 
  end while
end procedure

```

1.2.2 FIND-MAX

To find the maximum key in the heap H ,

- Simply return the item in the root
- Worst-case runtime: $\Theta(1)$



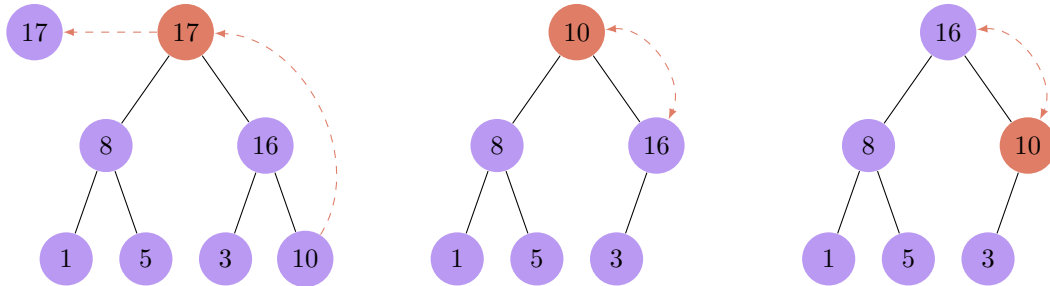
```

1: procedure FIND-MAX( $H$ )
2:   return  $H[1]$ 
3: end procedure

```

1.2.3 EXTRACT-MAX

- Save the item from the root in a temporary variable
- Replace the root with the rightmost item in the lowest level of the tree and decrement $H.heap\text{-}size$
- Repeatedly swap the item we moved with its largest child until the heap property is restored. This swapping process is called **bubble down**.
- Worst-case runtime: $\Theta(\lg n)$



```

1: procedure EXTRACT-MAX( $H$ )
2:    $max \leftarrow H[1]$ 
3:    $H[1] \leftarrow H[H.heap\text{-}size]$ 
4:    $H.heap\text{-}size \leftarrow H.heap\text{-}size - 1$ 
5:   MAX-HEAPIFY( $H, 1$ )
6:   return  $max$ 
7: end procedure

```

```

1: procedure MAX-HEAPIFY( $H, i$ )
2:    $l \leftarrow \text{LEFT}(i)$ 
3:    $r \leftarrow \text{RIGHT}(i)$ 
4:   if  $l \leq H.\text{heap-size}$  and  $H[l] > H[i]$  then
5:      $\text{largest} \leftarrow l$ 
6:   else
7:      $\text{largest} \leftarrow i$ 
8:   end if
9:   if  $r \leq H.\text{heap-size}$  and  $H[r] > H[\text{largest}]$  then
10:     $\text{largest} \leftarrow r$ 
11:  end if
12:  if  $\text{largest} \neq i$  then
13:    swap  $H[i]$  with  $H[\text{largest}]$ 
14:    MAX-HEAPIFY( $H, \text{largest}$ )
15:  end if
16: end procedure

```

1.2.4 BUILD-MAX-HEAP

- Takes an array A of length n and builds a max-heap H from it.
- Worst-case runtime: $\Theta(n)$

```

1: procedure BUILD-MAX-HEAP( $A$ )
2:    $H.\text{heap-size} \leftarrow A.\text{length}$ 
3:   for  $i = \lfloor \frac{A.\text{length}}{2} \rfloor$  downto 1 do
4:     MAX-HEAPIFY( $H, i$ )
5:   end for
6: end procedure

```

DICTIONARIES

Data

- A set S
- Each element x has a **unique** key $x.key$

Operations

- $\text{SEARCH}(S, k)$
Return x in S with $x.key = k$ (or NIL).
- $\text{INSERT}(S, x)$
Add x to S – if S contains y with $y.key = x.key$, then *replace* y with x .
- $\text{DELETE}(S, x)$ ^a
Remove element x from S .

^aIf we are given the key k instead of the element x , we could do $\text{DELETE}(S, \text{SEARCH}(S, k))$

	SEARCH	INSERT	DELETE [#]
unsorted array	n	n	1
sorted* array	$\lg n$	n	n
unsorted linked list	n	n	1
sorted* linked list	n	n	1
direct access table	1	1	1
hash table	n	n	n
binary search tree	height of the BST	height of the BST	height of the BST
balanced search tree	$\lg n$	$\lg n$	$\lg n$

*: these require the keys to be ordered

#: here we are only doing deletion (without having to search for x)

2.1 Binary Search Trees

A binary search tree is a binary tree with the *binary-search-tree property*:

Let x be a node in a binary search tree. If y is a node in the left subtree of x , then $y.key \leq x.key$. If y is a node in the right subtree of x , then $y.key \geq x.key$.

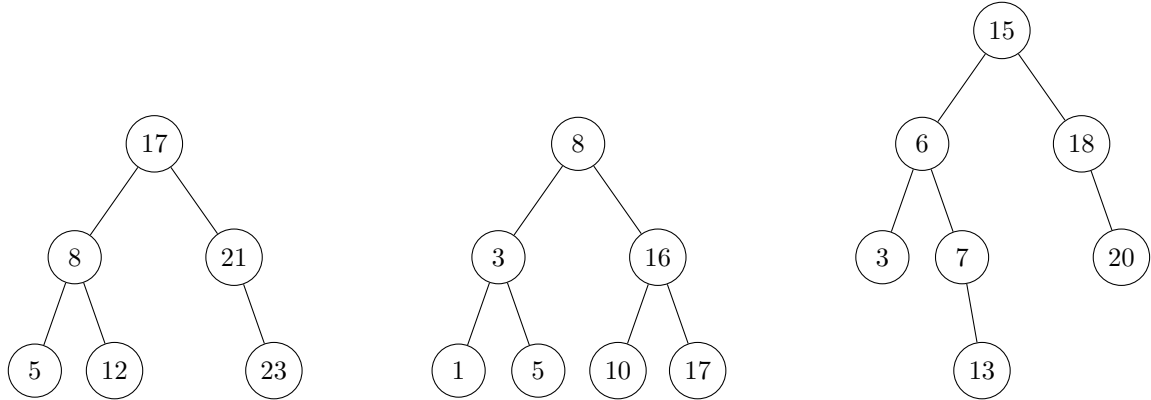


Figure 2.1: Examples of binary search trees

```
class Item:
    key: Any
    value: Any
```

```
class BST_Node:
    item: Item
    left: BST_Node
    right: BST_Node
```

```
class Dictionary:
    root: BST_Node
```

2.1.1 INSERT

```
1: procedure INSERT( $S, x$ )
2:    $S.root \leftarrow$  BST-INSERT( $S.root, x$ )
3: end procedure
```

```
1: procedure BST-INSERT( $root, x$ )
2:   # Insert  $x$  into stubtree at  $root$ ; return new
   root
3:   if  $root = \text{NIL}$  then
4:      $root \leftarrow$  BST_NODE( $x$ ) # Add  $x$ 
5:   else if  $x.key < root.item.key$  then
6:      $root.left \leftarrow$  BST-INSERT( $root.left, x$ )
7:   else if  $x.key > root.item.key$  then
8:      $root.right \leftarrow$  BST-INSERT( $root.right, x$ )
9:   else #  $x.key = root.item.key$ 
10:     $root.item \leftarrow x$  # replace with  $x$ 
11:   end if
12:   return  $root$ 
13: end procedure
```

2.1.2 SEARCH

```
1: procedure SEARCH( $S, k$ )
2:    $node \leftarrow$  BST-SEARCH( $S.root, k$ )
3:   if  $node = \text{NIL}$  then
4:     return NIL
5:   end if
6:   return  $node.item$ 
7: end procedure
```

```
1: procedure BST-SEARCH( $root, k$ )
2:   # Return node under root with key  $k$  (or NIL)
3:   if  $root = \text{NIL}$  then
4:     pass #  $k$  not in tree
5:   else if  $k < root.item.key$  then
6:      $root \leftarrow$  BST-SEARCH( $root.left, k$ )
7:   else if  $k > root.item.key$  then
8:      $root \leftarrow$  BST-SEARCH( $root.right, k$ )
9:   else #  $k = root.item.key$ 
10:    pass
11:   end if
12:   return  $root$ 
13: end procedure
```

2.1.3 DELETE

```
1: procedure DELETE( $S, k$ )
2:    $S.root \leftarrow$  BST-DELETE( $S.root, k$ )
3: end procedure
```

```
1: procedure BST-DELETE( $root, x$ )
2:   # Delete  $x$  from stubree at root; return new root
3:   if  $root = \text{NIL}$  then pass #  $x$  not in tree
4:   else if  $x < root.item.key$  then
5:      $root.left \leftarrow$  BST-DELETE( $root.left, x$ )
6:   else if  $x > root.item.key$  then
7:      $root.right \leftarrow$  BST-DELETE( $root.right, x$ )
8:   else #  $x.key = root.item.key$ 
9:     if  $root.left = \text{NIL}$  then
10:       $root \leftarrow root.right$  # could be NIL
11:     else if  $root.right = \text{NIL}$  then
12:       $root \leftarrow root.left$ 
13:     else # Replace  $root.item$  with its successor
14:       $root.item, root.right \leftarrow$  BST-DEL-MIN( $root.right$ )
15:     end if
16:   end if
17:   return  $root$ 
18: end procedure
```

```

1: procedure BST-DEL-MIN(root)
2:   # Remove element with smallest key under root; return item and root of resulting subtree
Require: root ≠ NIL
3:   if root.left = NIL then
4:     return root.item, root.right
5:   else
6:     item, root.left ← BST-DEL-MIN(root.left)
7:     return item, root
8:   end if
9: end procedure

```

2.2 Balanced Search trees

Despite the simplicity of the BST, it is not a very efficient data structure. The worst-case running time of the BST operations is proportional to the height of the tree, which is $\Theta(n)$, where n is the number of elements in the tree. The shape of a BST is determined by the order in which keys are inserted. If the keys are inserted in sorted order, the BST degenerates into a linked list.

We can improve the performance of the BST by making it more balanced. A *balanced BST* (also known as an *AVL tree* – Adelson-Velsky, Landis Tree) is one in which the heights of the two subtrees of any node differ by at most one. The height of a balanced BST is $\Theta(\log n)$, where n is the number of elements in the tree.

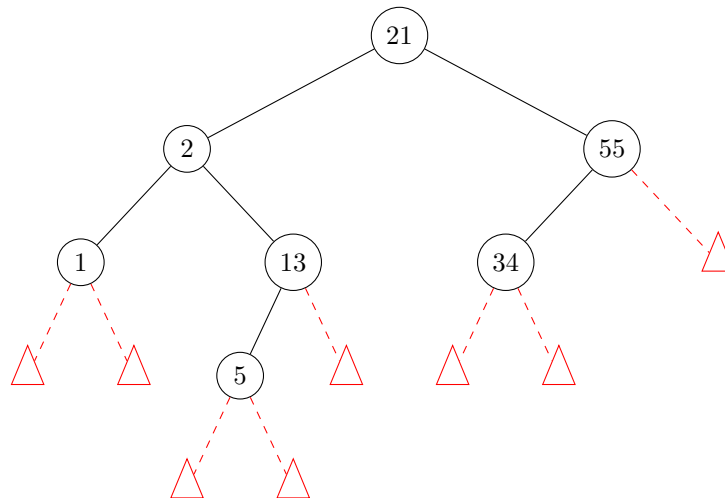


Figure 2.2: AVL balanced binary search tree

To implement an AVL tree, we need a mechanism to detect imbalance in the tree, and a way to restore balance. We will use the following definition of *balance factor* of a node x in a BST:

Definition 2.2.1 Balance Factor

An AVL balanced node x has a balance factor of -1 , 0 , or 1 . If the height of its left subtree is h_L , and the height of its right subtree is h_R , then x has a balance factor of $h_L - h_R$.

- If $h_R - h_L = 0$, then x is balanced.
- If $h_R - h_L = 1$, then x is right-heavy.
- If $h_R - h_L = -1$, then x is left-heavy.

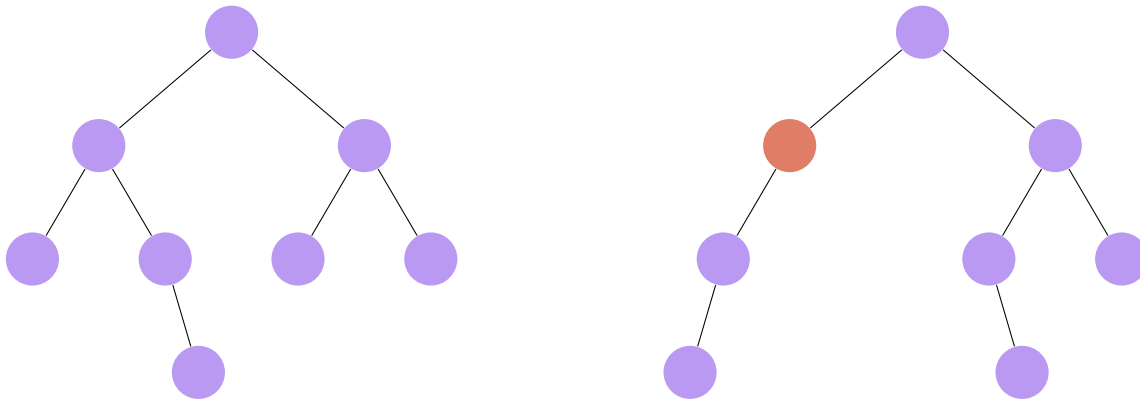


Figure 2.3: AVL balanced tree (left) and unbalanced tree (right)

Rotations

To restore balance, we need to perform a *rotation* on the tree. There are four types of rotations, depending on the balance factor of the node and its children. The following figure shows the four types of rotations.

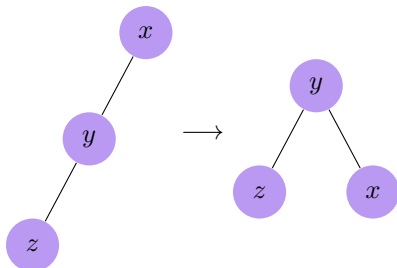


Figure 2.4: Single Left Rotation

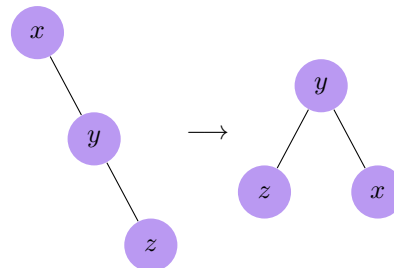


Figure 2.5: Single Left Rotation

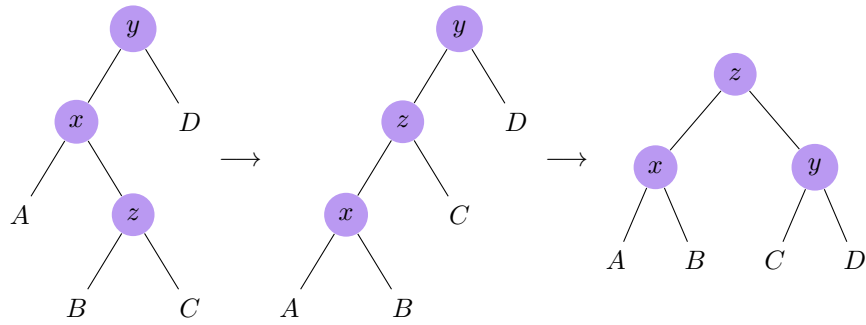


Figure 2.6: Double Left-Right Rotation

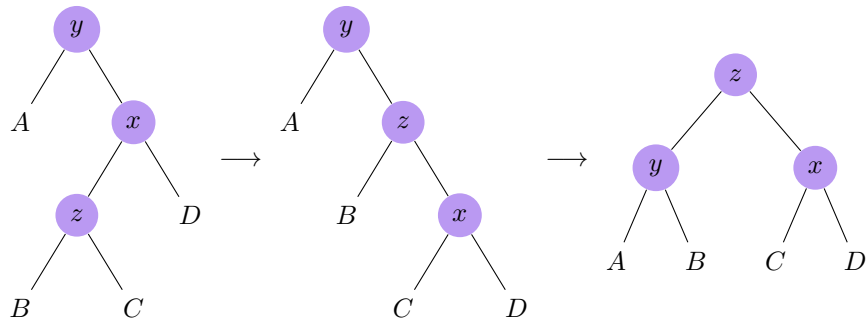


Figure 2.7: Double Right-Left Rotation

2.2.1 INSERT

```

1: procedure AVL-INSERT(root, x)
2:   # Insert x into the tree at root, return new root
3:   if root = NIL then
4:     root  $\leftarrow$  AVL_NODE(x) # add x
5:   else if x.key < root.item.key then
6:     root.left  $\leftarrow$  AVL-INSERT(root.left, x)
7:     root  $\leftarrow$  AVL-BALANCE-RIGHT(root)
8:   else if x.key > root.item.key then
9:     root.right  $\leftarrow$  AVL-INSERT(root.right, x)
10:    root  $\leftarrow$  AVL-BALANCE-LEFT(root)
11:  else # x.key = root.item.key
12:    root.item  $\leftarrow$  x # replace with x
13:  end if
14:  return root
15: end procedure

```

2.2.2 DELETE

```
procedure AVL-DELETE(root, x)
  # Delete x from the tree at root, return new root
  if root = NIL then
    pass # x not in tree
  else if x.key < root.item.key then
    root.left ← AVL-DELETE(root.left, x)
    root ← AVL-BALANCE-LEFT(root)
  else if x.key > root.item.key then
    root.right ← AVL-DELETE(root.right, x)
    root ← AVL-BALANCE-RIGHT(root)
  else # x.key = root.item.key
    if root.left = NIL then
      root ← root.right # could be NIL
    else if root.right = NIL then
      root ← root.left
    else
      if root.left.height > root.right.height then
        root.item, root.left ← AVL-DELETE-MAX(root.left)
      else
        root.item, root.right ← AVL-DELETE-MIN(root.right)
      end if
    end if
    root.height ← 1 + MAX(root.left.height, root.right.height)
  end if
  return root
end procedure
```

```
procedure AVL-DEL-MAX(root)
  # Delete the maximum item from the tree at root, return new root and deleted item
Require: root ≠ NIL
  if root.right = NIL then
    return root.item, root.left
  else
    item, root.right ← AVL-DELETE-MAX(root.right)
    root ← AVL-BALANCE-RIGHT(root)
    return item, root
  end if
end procedure
```

2.2.3 Rebalancing

```
1: procedure AVL-BALANCE-LEFT(root)
Require: root ≠ NIL
2:   # First, recalculate height
3:   root.height ← 1 + MAX(root.left.height, root.right.height)
4:   # Then, rebalance the left, if necessary
5:   if root.right.height > root.left.height + 1 then
6:     # Check for double rotation
7:     if root.right.left.height > root.right.right.height then
8:       root.right ← AVL-ROTATE-RIGHT(root.right)
9:     end if
10:    root ← AVL-ROTATE-LEFT(root)
11:  end if
12:  return root
13: end procedure
```

```
1: procedure AVL-ROTATE-LEFT(parent)
Require: parent ≠ NIL, parent.right ≠ NIL
2:   # Rearrange references
3:   child ← parent.right
4:   parent.right ← child.left
5:   child.left ← parent
6:   # Update heights; parent first because it is now deeper
7:   parent.height ← 1 + MAX(parent.left.height, parent.right.height)
8:   child.height ← 1 + MAX(child.left.height, child.right.height)
9:   # Return new parent
10:  return child
11: end procedure
```

2.3 Hashing

- Universe U
The set of all keys. We assume that $|U|$ is very large.
- Hash Table T
An array of fixed size m . Each location $T[i]$ is called a *bucket*.
- Hash Function h
The hash function $h : U \rightarrow \{0, 1, \dots, m-1\}$ maps each key in U to an index in $\{0, 1, \dots, m-1\}$. For each key $k \in U$, $h(k)$ is called the *home bucket* of k .
To access item with key k , examine $T[h(k)]$.

A hash table is an effective data structure for implementing dictionaries. Although SEARCH for an element in a hash table can take as long as searching for an element in a linked list – $\Theta(n)$ time in the worst case – in practice, hashing performs extremely well. Under reasonable assumptions, the average time to search for an element in a hash table is $\mathcal{O}(1)$.

2.3.1 Direct Access Tables

Direct addressing is a simple technique that works well when the universe U of keys is reasonably small. If U is small, then we can use an array T of size $|U|$ to implement a dictionary, called a *direct access table*. The key k is used as an index into T to access the item with key k .

2.3.2 Hash Tables

The downside of direct addressing is apparent: if the universe U is large or infinite, storing a table T of size $|U|$ is impractical, and the set K of keys *actually stored* may be so small relative to Y that most of the space allocated for T would be wasted. Instead, we use a hash table.

However, when $m \ll |U|$, collisions are unavoidable. A *collision* occurs when two keys k_1 and k_2 (with $k_1 \neq k_2$) are mapped to the same bucket $h(k_1) = h(k_2)$. There are two ways to handle collisions: *open addressing* and *closed addressing / chaining*.

Open Addressing

In open addressing, if $T[h(k)]$ is occupied, then we search for the next available location in T to store the item with key k . We call the original hash function h_1 the *primary hash function*, such that $h_1(k)$ is the home bucket of k . We use the *probe sequence* $h(k, i)$ to determine the bucket to try after i collisions.

- *Linear Probing*

$$h(k, i) = (h_1(k) + i) \bmod m$$

Note that long clusters of occupied buckets can occur.

- *Quadratic Probing*

$$h(k, i) = (h_1(k) + c_1 i + c_2 i^2) \bmod m$$

c_1 and c_2 are constants dependent on m .

- *Double Hashing*

$$h(k, i) = (h_1(k) + i \cdot (h_2(k))) \bmod m, \text{ where } h_2(k) \text{ is a secondary hash function.}$$

Close Addressing / Chaining

In close addressing, we use a linked list to store the items in each bucket. Each nonempty slot points to a linked list, and all the elements that hash to the same slot go into that slot's linked list.

The average-case performance of the hash table depends on how evenly the hash function h distributes the keys across the buckets in the table. The *simple uniform hashing assumption* (SUHA)

states that any given key is equally likely to hash into any of the m slots of the table, independently of where any other elements has hashed to. Under this assumption, the expected number of keys in each bucket is the same.

The expected number of keys in a bucket is $\frac{n}{m}$, where n is the number of items in the table and m is the size of the table. This ratio is called the *load factor* of the hash table, and we denote it by α .

DYNAMIC ARRAY

C styled arrays are static, meaning that they have a fixed size. In this chapter, we will learn how to implement a dynamic array, which is a data structure that can grow and shrink in size.

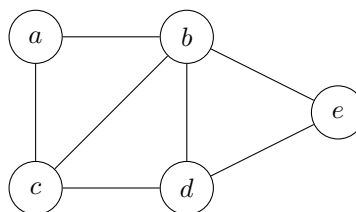
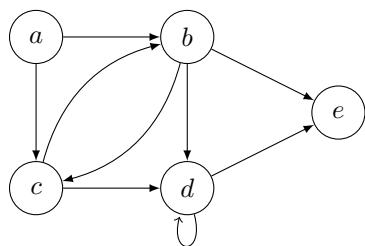
```
class DynamicArray {
    capacity: integer # room for elements
    size:      integer # actual number of elements
};
```

```
1: procedure INSERT( $A, x$ )
2:   if  $A.size = A.capacity$  then
3:      $A \leftarrow \text{RESIZE}(A)$ 
4:   end if
5:    $A.size \leftarrow A.size + 1$ 
6:    $A[A.size] \leftarrow x$ 
7: end procedure
```

```
1: procedure RESIZE( $A$ )
2:    $B \leftarrow \text{DYNAMICARRAY}(2 \times A.capacity)$ 
3:   for  $i = 1$  to  $A.size$  do
4:     INSERT( $B, A[i]$ )
5:   end for
6:   return  $B$ 
7: end procedure
```

To analyze the running time of the above algorithm, see this example using accounting method for amortized analysis. The amortized running time of the above algorithm is $\Theta(1)$.

GRAPHS



4.1 Graphs

4.1.1 Graphs

Define a graph $G = \{V, E\}$

Representations

- Adjacency matrix

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	1	1
c	1	1	0	1	0
d	0	1	1	0	1
e	0	1	0	1	0

Complexity: let $n = |V|$ and $m = |E|$

- Space: $O(n^2)$
- Edge query: $\Theta(1)$ time

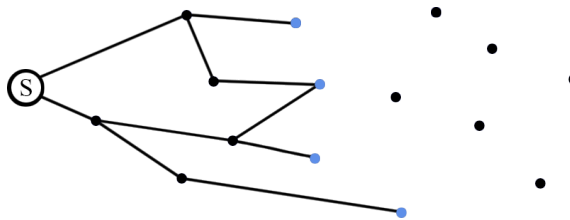
- Adjacency list

Complexity: let $n = |V|$ and $m = |E|$

- Space: $\Theta(n + m)$
- Edge query: $\Theta(n)$ in worst-case time

4.1.2 Breadth-First Search

In breadth first search, we start at a source $s \in V$, and explore every vertex reachable from a , using only edges.



We assign each vertex $v \in V$ a color, which can be one of the following:

- White: v has not been discovered
- Gray: v has been discovered, but not explored
- Black: v has been discovered and explored

Define $\pi[v]$ to be the predecessor of v in the breadth-first search tree.

Define $d[v]$ to be the distance from s to v in the breadth-first search tree.

We use a queue to keep track of the vertices that we have discovered, but not yet explored.

```

1: procedure BFS( $G, s$ )
2:   # Initialize tracking info for all vertices
3:   for  $v \in G.V$  do
4:     colour[ $v$ ] = white
5:      $\pi[v] \leftarrow \text{NIL}$ 
6:      $d[v] \leftarrow \infty$ 
7:   end for
8:   # Initialize empty queue and source vertex tracking info
9:    $Q \leftarrow \text{MAKE-QUEUE}()$ 
10:  colour[ $s$ ]  $\leftarrow$  gray
11:   $\pi[s] \leftarrow \text{NIL}$ 
12:   $d[s] \leftarrow 0$ 
13:  ENQUEUE( $Q, s$ )
14:  # Main loop. Loop Invariant:  $Q$  contains all (and only) grey vertices
15:  while  $Q \neq \text{EMPTY-QUEUE}$  do
16:     $u \leftarrow \text{DEQUEUE}()$ 
17:    for  $v \in G.Adj[u]$  do
18:      if colour[ $v$ ]  $\leftarrow$  white then
19:        colour[ $v$ ]  $\leftarrow$  gray
20:         $\pi[v] \leftarrow u$ 
21:         $d[v] \leftarrow d[u] + 1$ 
22:        ENQUEUE( $Q, v$ )
23:      end if
24:    end for
25:    colour[ $u$ ] = black
26:  end while
27: end procedure

```

In breath first search,

- Each vertex is enqueued at most once
- Each vertex is dequeued at most once
- Each adjacent list is examined at most once
- The time complexity is $\Theta(n + m)$

BFS Finds Shortest Paths

Define $\delta(s, v)$ to be the length of the shortest path from vertex s to vertex v (i.e. the smallest number of edges in any path from s to v). If there is no path from s to v , then $\delta(s, v) = \infty$. Note that this definition will change when we consider **weighted** graphs.

Theorem 4.1.1 Let $G = \{V, E\}$ be a graph, and let $s \in V$. Then, after BFS(G, s), $\forall v \in V, \delta(s, v) = d[v]$

To prove this theorem, we will need to prove the following lemmas first.

Lemma 4.1.1 $\forall (u, v) \in E, \delta(s, v) \leq \delta(s, u) + 1$

Proof. (idea)

If $\delta(s, u) = \infty$, then the claim holds trivially.

If $\delta(s, u) \neq \infty$, then u is reachable from s .

Thus, v is also reachable from s .

Thus, the shortest path from s to v is no longer than the shortest path from s to u , plus the edge (u, v) .

Hence, $\delta(s, v) \leq \delta(s, u) + 1$. ■

Lemma 4.1.2 At any point during BFS, $\forall v \in V, d[v] \geq \delta(s, v)$

Proof. (idea)

Use induction on the number of ENQUEUE operations.

Immediately after we do the first ENQUEUE operation, $d[s] = 0$, and $\delta(s, s) = 0$.

We also have $d[v] = \infty$, and $\delta(s, v) = \infty$ for all $v \in V - \{s\}$.

Now, consider some vertex v that is first discovered while visiting a vertex u .

By the IH, we have $d[u] \geq \delta(s, u)$.

Hence, $d[v] = d[u] + 1 \geq \delta(s, u) + 1$ by Lemma 4.2.1.

Then v is painted grey and $d[v]$ is not changed for the rest of the algorithm. ■

Lemma 4.1.3 If $Q = \langle v_1, \dots, v_r \rangle$, then $d[v_i] \leq d[v_{i+1}]$ for all $i \in \{1, \dots, r-1\}$ and $d[v_r] \leq d[v_1] + 1$

Proof. (sketch)

Use induction on the number of DEQUEUE / ENQUEUE operations.

When $Q = \langle s \rangle$, the claim holds trivially.

To prove the inductive step, we need to show that the lemma holds after applying DEQUEUE / ENQUEUE to $Q = \langle v_1, \dots, v_3 \rangle$.

- Case 1

If we perform a DEQUEUE operation, then $Q = \langle v_2, \dots, v_3 \rangle$ afterwards.

By the IH, $d[v_r] \leq d[v_1] + 1$ and $d[v_1] \leq d[v_2]$.

Hence, $d[v_r] \leq d[v_2] + 1$.

All other inequalities are unaffected.

- Case 2

If we perform an ENQUEUE operation, then $Q = \langle v_1, \dots, v_{r+1} \rangle$ afterwards.

We discover v_{r+1} while visiting some vertex u , so $d[v_{r+1}] = d[u] + 1$.

Vertex u must have been the previous vertex dequeued from the queue.

Hence, either v_1 was discovered while visiting u , in which case $[v_1] = d[u] + 1$, or Q was equal to $\langle u_2, v_1, \dots \rangle$ at some prior point, in which case $d[u] \leq d[v_1]$ by IH.

Hence, $d[v_{r+1}] = d[u] + 1 \leq d[v_1] + 1$.

Otherwise, $d[v_r] \leq d[u] + 1 = d[v_{r+1}]$ by the IH.

■

Now, we can prove the theorem.

Proof. To derive a contradiction, suppose $d[v] \neq \delta(s, v)$ for some vertex $v \in V$.

Suppose v is a vertex with minimal $\delta(s, v)$ for which this is satisfied.

By Lemma 4.1.2, we have $d[v] > \delta(s, v)$.

Because we chose v with minimal $\delta(s, v)$, we have $d[u] = \delta(s, u)$.

Hence, $d[v] > \delta(s, v) = \delta(s, u) + 1 = d[u] + 1$.

Consider the colour of v when we first dequeue u from Q .

- If v is painted **white**, then we set $d[v] = d[u] + 1$, which is a contradiction.
- If v is painted **black**, then v was in the queue before u . By Lemma 4.1.3, we have $d[v] \leq d[u]$, which is a contradiction.
- If v is painted **grey**, then v was discovered while visiting some vertex w that was dequeued earlier than u . Hence, $d[v] = d[w] + 1$, and by Lemma 4.1.3 we have $d[w] \leq d[u]$. So $d[v] \leq d[u] + 1$, which is a contradiction.

■

Time complexity of BFS

- Initialization (painting vertices white, setting entries of d to ∞ and entries of π to NIL) takes $\Theta(|V|)$ time.
- After initialization, we never paint a vertex white.
- Thus, each vertex is enqueued/dequeued at most once.
- Hence, we spend $\mathcal{O}(|V|)$ time doing queue operations.
- Every time we dequeue a vertex, we scan its out-neighbourhood to discover its neighbours:
 - With an **adjacency list**, we consider each edge at most once (or at most twice in an undirected graph), so in total we need at most $\mathcal{O}(|E|)$ time to consider all of the edges.
 - With an adjacency matrix, we scan each row of the matrix at most once, so in total we need at most $\mathcal{O}(|V|^2)$ time to consider all of the edges.

Using adjacency list: $\mathcal{O}(|V| + |E|)$

Using adjacency matrix: $\mathcal{O}(|V|^2)$

4.1.3 Depth-First Search

- In DFS, we walk through the graph as far as possible until we hit a dead end – when this happens, we backtrack to an undiscovered vertex.
- Similar to BFS, we paint vertices as we go:
 - Painted white: undiscovered
 - Painted grey: discovered but not yet visited
 - Painted black: visited
- Instead of storing distances/depths, we store timestamps:
 - $disc[v]$ = time at which v is first discovered
 - $vis[v]$ = time at which we finish visiting v
- One approach is to simply replace the *queue* from the BFS algorithm with a *stack*. This gives us an **iterative** DFS algorithm.
- However, it is more natural to write DFS as a recursive algorithm.

```
1: procedure DFS(G)
2:   # Initialization
3:   for each  $v \in G.V$  do
4:      $d[v] \leftarrow f[v] \leftarrow \infty$ 
5:      $\pi[v] \leftarrow \text{NIL}$ 
6:   end for
7:   time  $\leftarrow 0$  # global
8:   # Main loop
9:   for each  $v \in G.V$  do
10:    if  $d[v] = \infty$  then
11:      # colour[v] = white
12:      DFS-VISIT( $G, v$ )
13:    end if
14:  end for
15: end procedure
```

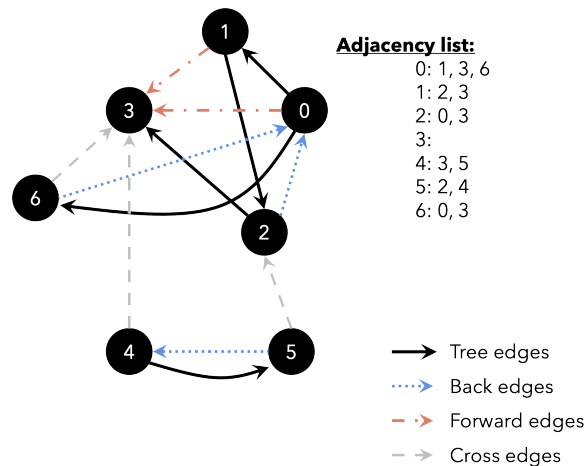
```
1: procedure DFS-VISIT( $G, v$ )
2:   # Discovered (colour[v] = grey)
3:    $d[v] \leftarrow \text{time} \leftarrow \text{time} + 1$ 
4:   # Do something with  $v$ , if desired
5:   # Explore  $v$ 's adjacency list
6:   for each  $u \in G.adj[v]$  do
7:     if  $d[u] = \infty$  then
8:       # colour[u] = white
9:        $\pi[u] = v$ 
10:      DFS-VISIT( $G, u$ )
11:     end if
12:   end for
13:   # Finished (colour[v] = black)
14:    $f[v] \leftarrow \text{time} \leftarrow \text{time} + 1$ 
15: end procedure
```

DFS Forests

We classify each edge (u, v) based on the colour of v when we consider this edge:

- *Tree edges* are the edges $u, v \in E$ that form the DFS forest stored by π
The vertex v is painted **white** when (u, v) is considered
- *Back edges* are the edges $(u, v) \in E$ such that v is an ancestor of u in the DFS forest
The vertex v is painted **grey** when (u, v) is considered

- **Forward edges** are the edges $(u, v) \in E$ such that v is a descendant of u in the DFS forest
The vertex v is painted **black** when (u, v) is considered
- **Cross edges** are all other edges $(u, v) \in E$ that are not part of the DFS forest (i.e. v is neither an ancestor nor a descendant of u)
The vertex v is painted **black** when (u, v) is considered



Note that for undirected graphs, there are NO forward edges and NO cross edges.

Time Complexity

The run-time of DFS is similar to BFS.

Using adjacency list: $\mathcal{O}(|V| + |E|)$

Using adjacency matrix: $\mathcal{O}(|V|^2)$

Parenthesis theorem

Theorem 4.1.2 After performing $\text{DFS}(G = (V, E))$, for any two vertices $u, v \in V$, exactly one of the following statements holds:

- 1 The intervals $[disc[u], vis[u]]$ and $[disc[v], vis[v]]$ are disjoint, and neither u nor v is a descendant of the other in the DFS forest.
- 2 The interval $[disc[u], vis[u]]$ is contained entirely in the interval $[disc[v], vis[v]]$, and u is a descendant of v in the DFS forest.
- 3 The interval $[disc[v], vis[v]]$ is contained entirely in the interval $[disc[u], vis[u]]$, and v is a descendant of u in the DFS forest.

Proof. (sketch)

Suppose that $disc[u] < disc[v]$.

- Case 1: $disc[v] < vis[u]$



v is first discovered while u is painted grey.

So v is a descendant of u .

We don't backtrack to u until we have finished visiting v .

Therefore, we paint v black and set $vis[v]$ before backtracking to u . Hence, $vis[v] < vis[u]$.

- Case 2: $vis[u] < disc[v]$



v is first discovered while u is painted black.

So v is not a descendant of u .

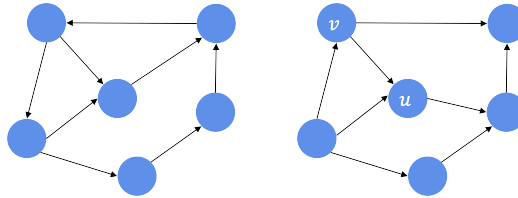
Since $disc[u] < disc[v]$, u is not a descendant of v .

Since $disc[u] < vis[u]$ and $disc[v] < vis[v]$, we have $disc[u] < vis[u] < disc[v] < vis[v]$.

When $disc[v] < disc[u]$, proof is symmetric. ■

4.1.4 Strongly connected

Recall that an undirected graph is connected if and only if there is a path of edges between any pair of vertices in V . What about directed graphs? A directed graph $G = (V, E)$ is strongly connected if and only if, for all $u, v \in V$, there is a path of edges from u to v in G .



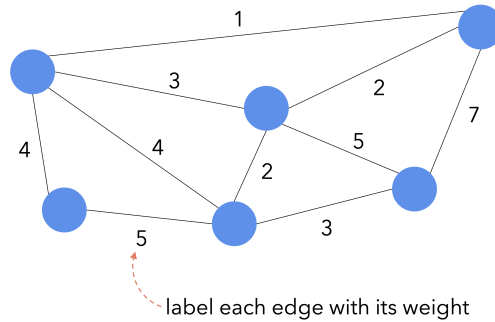
A strongly connected graph (left) and a non-strongly connected graph (right, where v is not reachable from u)

- Run DFS on the **transpose** of G (reverse edge direction of all edges in G)
- *Reorder* vertices in decreasing order of finish time (in $G.V$ and within each adjacency list)
Run *DFS* on G using the new ordering of vertices
- Each DFS tree is one strongly connected component

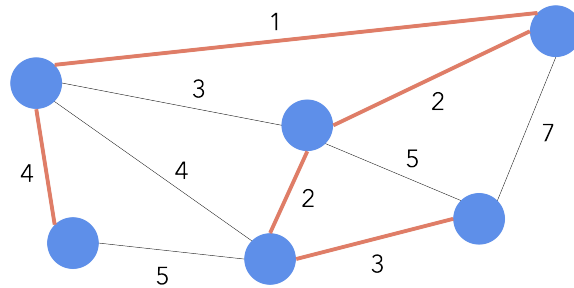
4.2

Minimum Spanning Trees

In a *weighted graph* $G = (V, E)$, each edge $e \in E$ has a weight $w(e)$



A *minimum spanning tree* (MST) T of a weighted, undirected, connected graph $G = (V, E)$ is an acyclic subset of E that connects all of the vertices in V and whose total weight $\sum_{\{u,v\} \in T} w(\{u,v\})$ is minimized.

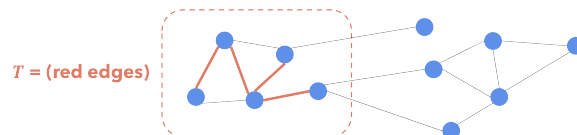


Above, we define an MST as a set of edges. We also often call the graph $M = (V, T)$ an MST of G .

The graph $M = (V, T)$ belongs to a special class of graphs called **trees**, which can be defined in many ways.

4.2.1 Prim's Algorithm

In Prim's Algorithm, we start with an empty tree T , add edges one at a time to T until we have an MST. At each step, pick the **smallest** edge that connects some vertex in T to a vertex outside T .



```

procedure PRIM( $G, w : E \rightarrow \mathbb{R}, r \in G.V$ )
  # Initialization
   $T = \emptyset$ 
   $Q = \text{MAKE-QUEUE}(G.V)$ 
  # Place all vertices in the queue
  for  $v \in G.V$  do
     $\text{pri}[v] = \infty$ 
     $\pi[v] = \text{NIL}$ 
     $Q.\text{INSERT}(v)$ 
  end for
  # Now, set  $r$  as the root
   $\text{pri}[r] = 0$ 
   $Q.\text{DECREASE-KEY}(r)$ 
  # Main loop
  while not  $Q.\text{EMPTY}()$  do
    # Connect a vertex with minimum priority (edge weight)
     $u = Q.\text{EXTRACT-MIN}()$ 
    if  $\pi[u] \neq \text{NIL}$  then  $T = T \cup \{\pi[u], u\}$ 
    end if
    # Update priorities of neighbors of  $u$ 
    for  $v \in G.\text{Adj}[u]$  do
      if  $v \in Q$  and  $w(u, v) < \text{pri}[v]$  then
         $\pi[v] = u$ 
         $\text{pri}[v] = w(u, v)$ 
         $Q.\text{DECREASE-KEY}(v)$ 
      end if
    end for
  end while
  return  $T$ 
end procedure

```

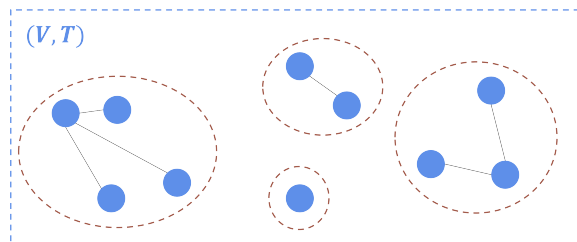
Runtime

- Initialization: $\Theta(n)$
- Main loop: n iterations; one EXTRACT-MIN each time
 $\Theta(n \lg n)$
- Inner loop: examine each edge twice; at most DECREASE-KEY per edge
 $\Theta(m \lg n)$
- Overall: $\Theta((m + n) \lg n)$

Note that to make DECREASE-KEY efficient, we need to track $\text{index}[v]$ – the position of v in the min-heap – and update the index values during each swap.

4.2.2 Kruskal's algorithm

In Kruskal's algorithm, we repeatedly pick the **cheapest edge** of the graph G that **does not create a cycle** in the current set of edges.



```
1: procedure KRUSKAL( $G, w$ )
2:   # Sort edges:  $w(e_1) \leq w(e_2) \leq \dots \leq w(e_m)$ 
3:    $T = \emptyset$ 
4:   for each  $e_i = e_1, e_2, \dots, e_m$  do
5:     let  $(u_i, v_i) = e_i$ 
6:     if  $u_i, v_i$  are not already connected by an edge in  $T$  then
7:        $T = T \cup \{e_i\}$ 
8:     end if
9:   end for
10:  return  $T$ 
11: end procedure
```

To determine if u_i, v_i are already connected, we can use BFS / DFS, which takes $\Theta(n)$ time¹. The overall run-time would be $\Theta(m \cdot n)$.

To make this algorithm more efficient, we can use disjoint sets.

¹BFS / DFS takes $\Theta(m + n)$ time. Since $m \leq n - 1$, this step takes $\Theta(n)$ time.

```

1: procedure KRUSKAL( $G, w$ )
2:   # Sort edges:  $w(e_1) \leq w(e_2) \leq \dots \leq w(e_m)$ 
3:    $T = \emptyset$ 
4:   for  $do v \in G.V$ 
5:     MAKE-SET( $v$ )
6:   end for
7:   for each  $e_i = e_1, e_2, \dots, e_m$  do
8:     let  $(u_i, v_i) = e_i$ 
9:     if FIND-SET( $u_i$ )  $\neq$  FIND-SET( $v_i$ ) then
10:       $T = T \cup \{e_i\}$ 
11:      UNION( $u_i, v_i$ )
12:    end if
13:  end for
14:  return  $T$ 
15: end procedure

```

4.3

Disjoint Sets

Data

- A collection $\{S, \dots, S\}$ of disjoint, dynamic sets where $S_i \neq \emptyset$
Each set in this collection contains a single *representative*

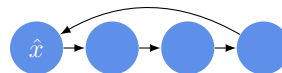
Operations

- MAKE-SET(x)
Precondition: $x \in U, x \notin S_1, \dots, x \notin S_n$
Create a new set $S_x = \{x\}$, add S_x to the collection
- FIND-SET(x)
Return the representative element for set S
s.t. $x \in S$, or NIL if x no such S exists

To check if two elements belong to the same set, we can simply check if they have the same representative. Note that we cannot ask for all the elements of a given set, or for the number of elements in a set.

4.3.1 Implementation

Implementation 1: Circular Linked List, Boolean *rep*



We use a *.rep* field to indicate whether a node is a representative. *.rep* is set to **true** for the representative of each set, and **false** for all other nodes.

- MAKE-SET(x): $\Theta(1)$

We only need to create a linked list with a single node

- FIND-SET(x): $\Theta(n)$

We need to traverse all list nodes to find x and then the representative

- UNION(x, y): $\Theta(n_1 + n_2)$

We need to first identify the two representatives.²

- If x and y share the same representative, do nothing;
- Otherwise, merge the two circular linked lists via changing the *next* attribute for two nodes (conveniently, we choose the representatives, but any two nodes would work), then pick a new representative for the merged set.

We should consider the total amortized complexity of m operations.

- Upper Bound

Given m operations, we know that each list has size $\leq m$. Then, each operations takes time $\mathcal{O}(m)$. In total, the amortized complexity is $\mathcal{O}(m^2)$.

- Lower Bound

Consider $\frac{m}{3}$ MAKE-SET operations, $\frac{m}{3}$ UNION operations, and $\frac{m}{3}$ FIND-SET operations.

The MAKE-SET and UNION operations take $\Omega(m)$ time. The FIND-SET operations take $\Omega\left(\frac{m^2}{9}\right)$ time.

Thus, the total time is $\Omega(m^2)$.

Thus, we can conclude that the amortized complexity of m operations is $\Theta(m^2)$.

Implementation 2: Circular Linked List, Pointer to *rep*

This is similar to implementation 1, except that we use a pointer to the representative instead of a boolean *rep* field.

Now, FIND-SET takes $\Theta(1)$ time, and UNION takes $\Theta(n_1)$ time since the *.rep* field needs to be updates for one of the to-be-merged sets.

- Upper Bound

This is the same as implementation 1.

- Lower Bound

Consider $\frac{m}{2}$ MAKE-SET operations and $\frac{m}{2}$ UNION operations.

Then UNION operations would take $\sum_{i=1}^{\frac{m}{2}-1} i \in \Omega(m^2)$

²If the representatives are given, UNION(x, y) is $\Theta(1)$.

Implementation 3: Circular Linked List, Pointer to *rep*, with *size*

In this implementation, we add a *size* field to *rep* to keep track of the number of elements in the set. During UNION, we merge the smaller set into the larger set, so there will be less updates to the *rep* field. We then update the *size* field for the representative of the merged set.

- Upper Bound

For any sequence of m operations, let n be the number of MAKE-SET operations. Then, the total time for all UNION operations is less than or equal to the total number of updates to the *rep* field, for all nodes.

For each element x , $x.rep$ is updated exactly when the set that contains x is smaller (or the same size) than the other set. Thus, this set must at least double in size after the UNION operation. Since there are n MAKE-SET operations, there can be at most $\log n$ updates to the *rep* attribute for every set.

In total, the amortized complexity is in $\mathcal{O}(n \log n + m) \in \mathbb{J} \log \setminus$.

- Lower Bound

Implementation 4: Inverted Trees

Each node has *parent* attribute, and representative is the root of the tree ($root.parent = root$). These trees do not have to be binary trees.

- MAKE-SET(x): $\Theta(1)$

We only need to create a tree with a single node

- FIND-SET(x): $\Theta(n)$

We need to follow *parent* pointers to find the representative. The time depends on how deep the chain / tree is.

- UNION(x, y): $\Theta(n)$

We need to find the two representatives first via FIND-SET, change one of their *parent* pointer to point to the other representative (one more level is added to the chain/tree).

Overall, the total worst case of m operations is $\Theta(m^2)$, as it is possible for us to generate a single chain.

Implementation 5: Inverted Trees, with *size*

We are using the same *union by weight* idea as implementation 3, Overall, the total worst case running time is $\Theta(m \log m)$.

Implementation 6: Inverted Trees Path Compression

The problem is not the total size but rather the height, we want to keep things as close to the root as possible.

FIND-SET(x) would follow the pointers, as before, but also add constant-time work to edit the nodes it visits on its way to all point to the root directly as their *parent* attribute. Then, we just need to perform FIND-SET(x) on the bottom of every long chain formed to compress.

Implementation 7: Inverted Trees Path Compression, with *rank*

This is a “*union-by-rank*” idea where rank is an upper bound on height. For MAKE-SET and UNION, the rank can be treated as the same as the height. However, due to path compression, the height may get smaller (we don’t want to keep track of/compute that).

During UNION, make the root with smaller rank a child of the other tree.

- If $r_2 < r_1$, make tree with rank r_2 a child of tree with rank r_1 . r_1 does not change.
- If $r_1 = r_2$, make tree with rank r_2 a child of tree with rank r_1 , and increment r_1 by 1.

Overall, the total worst case run-time of m operations is $\Theta(m \log^* m)$, where $\log^* m$ is the iterated logarithm³.

³Intuitively, the iterated logarithm represents the number of times the logarithm function must be iteratively applied before the result is less than or equal to 1. For example, $\log^* 2 = 1$, $\log^* 2^2 = \log^* 4 = 1$, $\log^* 16 = 3$, $\log^* 2^{16} = 1 + \log^* 16 = 4$, $\log^* 2^{65536} = 1 + \log^* 65536 = 1 + \log^* 2^{16} = 5$

Part II

Algorithms

SORTING

5.1 Heap Sort

Implementation 1:

- Run BUILD-MAX-HEAP(A)
- Run EXTRACT-MAX(A) for $n-1$ times

Implementation 2:

- We modify the EXTRACT-MAX algorithm to swap the root with the last item

5.2 Quick Sort

```
1: procedure QUICKSORT( $A$ )
2:   if  $\text{LEN}(A) \leq 1$ 
3:     return  $A$ 
4:   end if
5:    $L, p, G \leftarrow \text{PARTITION}(A)$ 
6:   return QUICKSORT( $L$ ) +  $[p]$  + QUICKSORT( $G$ )
7: end procedure
```

5.2.1 Deterministic Quick Sort

In deterministic quick sort, we choose the pivot to be a certain index in the array. We can choose the pivot to be the first element, the last element, or the middle element.

- Run-time depends on input ordering
- Bad ordering would yield bad run-time, while random ordering would generally yield better run-time

5.2.2 Randomized Quick Sort

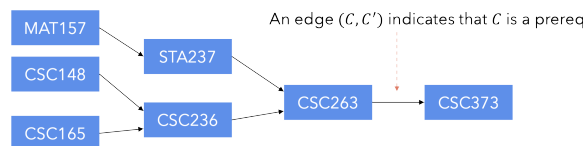
In randomized quick sort, we choose the pivot to be a random index in the array. Intuitively, the expected run-time would be the same as deterministic quick sort – $\Theta(n \lg n)$ – but the worst case run-time would be much better.

5.3

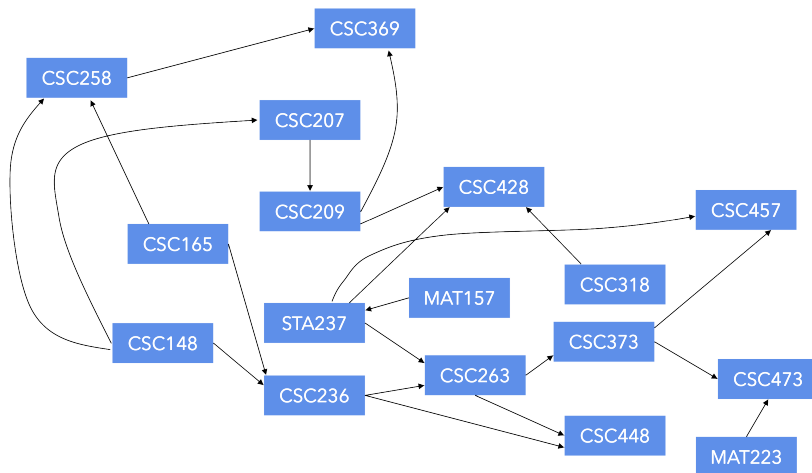
Topology Sort

Example. We wish to determine the sequence of courses to take during our undergraduate program

- A university course C may have a set of prerequisite courses that must be completed before C
- How can we determine a valid sequence of courses to take?

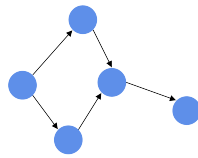


When we add more nodes/edges, things become less clear

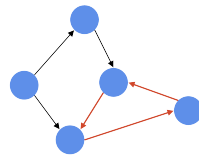


5.3.1 Directed Acyclic Graphs

A directed acyclic graph (or DAG) is a directed graph with no cycles



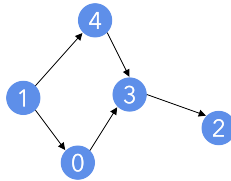
this **is** a DAG



this is **not** a DAG

5.3.2 Topological Sort

A topological sort/ordering of a DAG $G = (V, E)$ is a sequence of its vertices such that, if $(u, v) \in E$, then u appears before v in the sequence



1, 0, 4, 3, 2 is a topological ordering of these vertices

Given a DAG $G = (V, E)$, how can we efficiently compute a topological ordering of G ?

- Claim: if we perform DFS on a DAG G , the resulting DFS forest contains no back edges.
- Claim: every out-neighbour of every vertex v has an earlier ‘visited’ time than v .
 - If (v, u) is a tree / forward edge, then u is a descendant of v . By the Parenthesis Theorem, $[disc[u], vis[u]]$ is contained in $[disc[v], vis[v]]$. Hence, $vis[u] < vis[v]$.
 - If (v, u) is a cross edge, then u was already visited by the time we finish visiting v . Hence, $vis[u] < vis[v]$.

We claim that a directed graph G can be topologically sorted iff G contains no cycles iff DFS finds no back edges.

- 1 Maintain a linked list L while performing DFS on $G = (V, E)$
- 2 When we finish visiting a node v (by setting $vis[v]$ and painting v black), add v to the head of the linked list L
- 3 Once DFS terminates, L contains a list of all vertices in V sorted in decreasing order by their ‘visited’ times

Part III

Analysis

AVERAGE CASE ANALYSIS

In *average case analysis*, we are interested in the average performance of an algorithm. we take the average, or the expected value, over the distribution of the possible inputs.

For each n , define $S_n = \{\text{all inputs of size } n\}$, and if we consider the inputs to be random, then S_n is the sample space. For each $x \in S_n$, define $P(x)$ to be the probability that x will be chosen as the input. Define $t(x)$ as the number of steps performed on input x . t is the random variable.

Then, the average case running time is defined as

$$\begin{aligned} T(n) &= E[t] \\ &= \sum_{x \in S_n} P(x) \cdot t(x) \end{aligned}$$

Example. Consider the linear search algorithm on a linked list L .

```

1: procedure LINSEARCH( $L, x$ )
2:    $z \leftarrow L.head$ 
3:   while  $z \neq \text{NIL}$  and  $z.data \neq x$  do
4:      $z \leftarrow z.next$ 
5:   end while
6:   return  $z$ 
7: end procedure

```

Let S_n be the sample space of all linked lists of size n . Let $P(x)$ be the probability that x is chosen as the input. Let $t(x)$ be the number of steps performed on input x .

- We need to know S_n with probability

Consider the inputs $\text{input}_1 = ([1, 2, 3], 2)$ and $\text{input}_2 = ([\text{"a"}, \text{"b"}, \text{"c"}], \text{"b"})$, note that they will take the same steps. We only need one input for each possible value of t .

Define $S_n = \{([1, 2, \dots, n], 1), ([1, 2, \dots, n], 2), \dots, ([1, 2, \dots, n], n), ([1, 2, \dots, n], 0)\}$.

- We assume all the inputs happen equally likely, then $P(x) = \frac{1}{n+1}$.
- We need an exact formula for $t(x)$.

In practice, we choose some “key operations” s.t. counting **only** these operations is within a constant factor of total time – then set $t(x) = \text{number of key operations}$.

Here, we choose line 3, $z.data \neq x$, as the key operation.

$$\begin{aligned}
 \text{Then, we have } T(n) &= \sum_{(L,i) \in S_n} t(L,i) \cdot P(L,i) \\
 &= \frac{1}{n+1} \sum_{i=0}^n t([1, 2, \dots, n], i) \\
 &= \frac{1}{n+1} \left(t([1, 2, \dots, n], 0) + \sum_{i=1}^n i \right) \\
 &= \frac{1}{n+1} \left(n + \frac{n(n+1)}{2} \right) \\
 &= \frac{n}{n+1} + \frac{n}{2}
 \end{aligned}$$

Example. Consider $\text{SEARCH}(T, k)$ on a hash table T for a key k .

Assume t has m slots, and uses chaining to resolve collisions. Assume that prior to applying the `textscSearch` algorithm, the hash table contains n keys.

Assume the key k is samples uniformly at random from U .

Let $N(k)$ be the number of keys examined during search for k . $N(k)$ is the key operation.

$$\begin{aligned}
 E[N(k)] &= \sum_{k \in U} P[k] \cdot N(k) \\
 &= \sum_{i=0}^{m-1} \sum_{\substack{k \in U \\ h(k)=i}} P[k] \cdot N(k) && \text{regroup terms} \\
 &\leq \sum_{i=0}^{m-1} \sum_{\substack{k \in U \\ h(k)=i}} P[k] \cdot L_i && \text{since } N(k) \leq L_i \text{ when } h(k) = i \\
 &= \sum_{i=0}^{m-1} L_i \cdot \sum_{\substack{k \in U \\ h(k)=i}} P[k] \\
 &= \sum_{i=0}^{m-1} L_i \cdot P[h(k) = i]
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^{m-1} L_i \cdot \frac{1}{m} \\
&= \frac{1}{m} \sum_{i=0}^{m-1} L_i \\
&= \frac{n}{m}
\end{aligned}$$

Example. Consider QUICKSORT on an array A of size n .

Let $S_n = \{\text{all permutations of } [1, 2, \text{dots}, n]\}$

We assume an uniform distribution of the inputs, then $P(x) = \frac{1}{n!}$.

Let the random variable $T(A)$ be the total number of comparisons between elements of A .

Define $X_{i,j} = \begin{cases} 1 & \text{if } i \text{ is compared to } j \\ 0 & \text{otherwise} \end{cases}$ for $1 \leq i < j \leq n$.

Then, $T(A) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n X_{i,j}$.

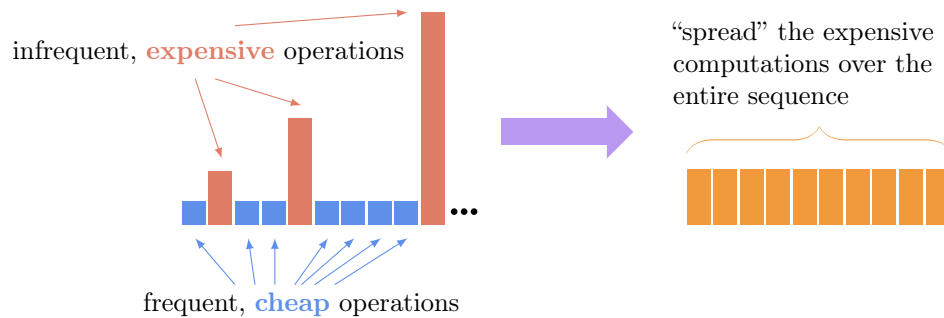
The probability $P(X_{i,j} = 1) = P(i \text{ or } j \text{ appear in } A \text{ before all other values in range } [i \dots j])$

$$\begin{aligned}
&= \frac{1}{j-i+1} + \frac{1}{j-i+1} \\
&= \frac{2}{j-i+1}
\end{aligned}$$

$$\begin{aligned}
\text{Then, } E[T(A)] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^n E[X_{i,j}] \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^n P(X_{i,j} = 1) \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{j-i+1} \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{\alpha+1} \quad \text{substitute } j-i \text{ with } \alpha \\
&< \sum_{i=1}^{n-1} \sum_{j=i+1}^n \frac{2}{k} \\
&= \sum_{i=1}^{n-1} \Theta(\lg n) \\
&= \Theta(n \lg n)
\end{aligned}$$

AMORTIZED ANALYSIS

In amortized analysis, we analyze the cost of a sequence of operations, not just a single operation. The amortized cost of a sequence of operations is the average cost per operation.



- The *worst-case sequence complexity* of a sequence S of k operations is the maximum possible total steps performed by S (taken over all possible inputs to the operations in S).
- The worst-case sequence complexity is **at most** $k \times$ the worst-case complexity of any individual operation in S .
- Suppose that the worst-case sequence complexity of a sequence of k operations is $T(k)$. Then the (worst-case) *amortized complexity* per operation of this sequence is $\frac{T(k)}{k}$.
- With amortized analysis, we take the average of the costs of multiple operations (as opposed to average-case analysis, where we calculate the cost of a single operation by averaging over the input distribution)

7.1

Aggregated Method

In the *aggregated method*, we determine the upper bound $T(n)$ on the total cost of a sequence of N operations, then calculate the average cost per operation as $\frac{T(n)}{n}$.

Example. Consider when we insert into an array. We increase the size of the array by 4 when it is $\frac{3}{4}$ full.

For simplicity, suppose that each insert with no resizing requires c steps, for some constant $c \in \mathbb{N}^+$ (so each insert with resizing requires $c \cdot n + c$ steps, where n is the number of items in the array prior to resizing).

Operation Number	Cost
1	c
2	c
3	c
4	$4c$
5	c
6	c
7	c
8	$7c$
9	c
10	$10c$
\vdots	\vdots

Within a sequence of k insertions, we need to resize $\lfloor \frac{k-1}{3} \rfloor$ times. For all k insert operations, the total cost is

$$T(k) = c \cdot \sum_{i=1}^{\lfloor \frac{k-1}{3} \rfloor} (3i+1) + c \cdot \left(k - \left\lfloor \frac{k-1}{3} \right\rfloor \right)$$

$$\begin{aligned}
 \text{Note that } c \cdot \sum_{i=1}^{\lfloor \frac{k-1}{3} \rfloor} (3i+1) &= c \cdot \sum_{i=1}^{\lfloor \frac{k-1}{3} \rfloor} 3i + \sum_{i=1}^{\lfloor \frac{k-1}{3} \rfloor} 1 \\
 &= 3c \cdot \sum_{i=1}^{\lfloor \frac{k-1}{3} \rfloor} i + c \cdot \left\lfloor \frac{k-1}{3} \right\rfloor \\
 &= \frac{3}{2}c \cdot \left\lfloor \frac{k-1}{3} \right\rfloor \cdot \left(\left\lfloor \frac{k-1}{3} \right\rfloor + 1 \right) + c \cdot \left\lfloor \frac{k-1}{3} \right\rfloor \in \Theta(k^2)
 \end{aligned}$$

Thus, $T(k)$ is $\Theta(k^2)$. The amortized cost per operation is $\frac{T(k)}{k} = \Theta(k)$.

Example. Consider a k digit binary counter.

In this problem, we count the total number of bits changed in the counter. We can do this by counting the number of times each bit changes.

Bit Number	Number of Changes
0	m
1	$\approx \frac{m}{2}$
2	$\approx \frac{m}{4}$
\vdots	\vdots
i	$\approx \frac{m}{2^i}$
\vdots	\vdots

$$\begin{aligned}
\text{Then, } T &= \sum_{i=0}^{(\lg m)-1} \frac{m}{2^i} \\
&= m \cdot \sum_{i=0}^{(\lg m)} \frac{1}{2^i} \\
&< m \sum_{i=0}^{\infty} \frac{1}{2^i} \\
&= 2m
\end{aligned}$$

7.2 Accounting Method

The *accounting method* is a form of aggregate analysis which assigns to each operation an amortized cost which may differ from its actual cost. Early operations have an amortized cost higher than their actual cost, which accumulates a saved “credit” that pays for later operations having an amortized cost lower than their actual cost. Because the credit begins at zero, the actual cost of a sequence of operations equals the amortized cost minus the accumulated credit. Because the credit is required to be non-negative, the amortized cost is an upper bound on the actual cost. Usually, many short-running operations accumulate such credit in small increments, while rare long-running operations decrease it drastically.

Example. Consider a sequence of m INSERT for dynamic array.

Recall that the “cost” is the actual run-time, while the “charge” is the estimated amortized time.

Note that for $k = 2^n + 1$, we need $2^n = k - 1$ for reading and $2^n + 1 = k$ for writing. Thus, the cost is $2k + 1$.

$$\text{Then, we know that } \text{cost}(\text{INSERT}(k)) = \begin{cases} 2k + 1 & \text{if } k = 2^n + 1 \\ 1 & \text{otherwise} \end{cases}$$

Define $\text{charge}(\text{INSERT}(k)) = \5 . We need \$1 for writing the new element, and save \$4 as credit.

We need to prove our credit invariant: every element in the second half of the array has \$4 credit.

Proof. Proof by induction on the number of operations done.

- init: 0 elements, 0 credits
- Consider one INSERT Assuming credit invariant holds

- If the array does not grow, then the new element gets \$4 credit. The credit invariant holds.
- If the array does grow, it must be full, the total credits is $\$4 \cdot \frac{n}{2} = \$2n$, enough to cover copying n elements. The new element gets \$4 credit. The credit invariant holds.

■

$$\begin{aligned}
 \text{Thus, Amortized} &\leq \frac{\text{WCSC}}{m} = \frac{\text{total cost}}{m} \\
 &\leq \frac{\text{total charge}}{m} \quad \text{because of credit invariant} \\
 &= \frac{5m}{m} \\
 &= 5
 \end{aligned}$$

LOWER BOUNDS

For upper bound $U(n)$ on problem complexity, we just need ONE algorithm with worst-case time $\Theta(U(n))$. For instance, the sorting has problem complexity $\mathcal{O}(n \log n)$ because heap sort takes time $\Theta(n \log n)$.

For lower bound $L(n)$ on problem complexity, we need a general argument that **every** algorithm takes worst-case time $\Omega(L(n))$. The goal is to prove that every algorithm that solves problem P requires time $\Omega(g(n))$ for some function $g(n)$. In contrast to upper bounds, we want our lower bounds $g(n)$ to be as large as possible.

Our algorithm is optimal with respect to worst-case time complexity when the bounds match. Lower bounds are usually proven only for a certain class of algorithms.

8.1

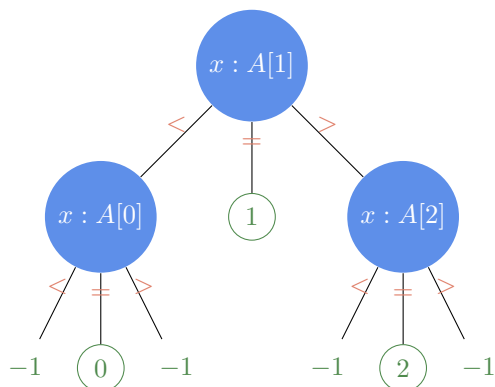
Comparison-Based Algorithms

A comparison-based algorithm is an algorithm that uses comparisons (for example, $>$, \geq , $<$, \leq , $=$, \neq , 3-way comparisons) on the input values with each other to solve a problem. For example, sorting is a comparison-based algorithm because it uses comparisons to determine the order of elements. These algorithms can be represented in the form of *comparison trees*.

It is conventional to use the same comparison operator in every node when possible. A single comparison tree can only represent the behaviour of an algorithm for a fixed input size. Each input size corresponds to a different tree (with a common structure). Then, an algorithm would lead to a family of comparison trees (one for each input size). For any specific input, the execution of the algorithm can be traced in the decision tree as a path from the root to a leaf node.

- Internal nodes (nodes with children) represent comparisons made by the algorithm: the root is the first comparison, its children the next comparison (depends on the result of the first), and so on.
- Edges show the outcome/result of each comparison

- External nodes (leaves) show the output of the algorithm (different nodes can have the same value since the algorithm can arrive at the same output in different ways).



The number of comparisons performed by the algorithm on any input is equal to the length of its execution path.

- The worst-case time complexity is proportional to the height of the tree
- The average-case time complexity is proportional to the average depth of the leaves

For any problem that has at least K different outcomes, every decision tree for the algorithm contains at least K leaves (at least 1 leaf for each outcome). In any tree with K leaves and fixed branching factor has height $\Omega(\log K)$.

Example. Consider searching a sorted list.

Let P_1 be such problem.

- Input: a sorted list A , a value x
- Output: index i such that $A[i] = x$ or -1 if x is not in A

Every comparison tree has at least 1 leaf for each outcome ($\geq n + 1$ since n elements and x could be not in the list)

Then, the worst-case run-time is $\Omega(\text{height}) = \Omega(\log(\text{number of leaves})) = \Omega(\log(n + 1))$.

Example. Consider sorting a list.

Let P_2 be such problem.

- Input: a list A of n elements
- Outcome: permutation of $[0, 1, \dots, n - 1]$, such that $A[i]$ represents the position of $A[i]$ in the sorted list

The number of leaves \geq the number of permutations $= n!$.

The height of the comparison tree $\geq \log(n!)$.

Then, the worst-case run-time is $\Omega(\log(n!)) = \Omega(n \log n)$.

8.2

Information Theory Lower Bounds

worst-case runtime = height of comparison tree

$$\begin{aligned} &\geq \log(\text{number of leaves}) \\ &\geq \log(\text{number of possible outcomes}) \\ &\geq \log(\text{number of outputs}) \end{aligned}$$

Example. Consider searching an unsorted list.

Let P_3 be such problem.

- Input: an unsorted list A , a value x
- Output: index i such that $A[i] = x$ or -1 if x is not in A

By applying information theory lower bound, since the comparison tree has at least $n + 1$ leaves (one for each possible output), we have worst case runtime $\geq \log(n + 1)$

8.3

Adversary / Adversarial Arguments

The goal is to show that every correct algorithm has a worst-case running time $\geq L(n)$. In practice, we often show the contrapositive: if there is an algorithm with running time $\leq L(n)$, then the algorithm is incorrect.

runtime = count key operations to use correct, exact lower bounds.

Example. Consider searching an unsorted list.

We count the number of comparisons ($x \stackrel{?}{=} A[j]$)

Try $L(n) = n$ because we know how to solve within n comparisons. We prove all algorithms that makes $< n$ comparisons are incorrect.

Consider one such algorithm. We work through the comparison tree.

- For each comparison $x \stackrel{?}{=} A[j]$, set $A[j] \neq x$, go down the **False** branch
- When we reach the leaf, there are at most $n - 1$ elements set in A . There is at least one element $A[i]$ whose value is not set.
 - If the algorithm returns i , let $A[i] = 0 \neq x$
 - If the algorithm returns -1 , let $A[i] = 1 = x$

In either case, this algorithm is incorrect.

Example. Consider Finding the maximum element in an unsorted list.

- Input: an unsorted list A of n elements
- Output: index i such that $A[i] \geq A[j]$ for all $j \in \{0, 1, \dots, n - 1\}$

By applying the information theory lower bound, we have worst-case runtime $\geq \Omega(\log(n))$

For the adversary argument, the algorithm requires less than $n - 1$ comparisons to find the maximum element.

We start at the root and traverse the comparison tree.

- Every comparison $A[i] \stackrel{?}{<} A[j]$ eliminates one possible location for the max.
- There is at most $n - 2$ comparisons, so at most $n - 2$ positions eliminated. There is at least two potential positions for the maximum

No matter which index is output, there is at least one index that makes it incorrect.

Part IV

Appendices

BIBLIOGRAPHY

Books

INDEX

- Amortized Analysis, 55
 - Accounting Method, 57
 - Aggregated Method, 56
 - Amortized Complexity, 55
 - Worst-Case Sequence Complexity, 55
- Average Case Analysis, 51
- AVL Tree, 18
 - Balance Factor, 19
 - Rotation, 19
- Balance Factor, 18
- Balanced Binary Search Tree, 18
- Binary Search Tree, 16
 - Binary-Search-Tree Property, 16
 - linear probing, 23
 - quadratic probing, 23
 - primary hash function, 23
 - Probe Sequence, 23
 - Simple Uniform Hashing Assumption, 23
 - Universe, 22
- Heap, 10
- Minimum Spanning Tree, 35
- Parenthesis Theorem, 33
- Prim's Algorithm, 35
- DFS Tree
 - Back Edge, 32
 - Cross Edge, 33
 - Forward Edge, 33
- Disjoint Sets, 38
- Graph
 - Weighted Graph, 35
- Hash Table, 22
 - bucket, 22
 - Closed Addressing / Chaining, 23
 - collision, 23
 - Direct Access Table, 23
 - hash function, 22
 - home bucket, 22
 - load factor, 24
 - Open Addressing, 23
 - open addressing
 - double hashing, 23