

Crash Course Coding Companion

SAMUEL HSIANG
THOMAS JEFFERSON HIGH SCHOOL FOR SCIENCE AND TECHNOLOGY
`samuel.c.hsiang@gmail.com`

August 7, 2015

Copyright, publishing, and licensing information should go here. I don't have anything fancy like that. It is of my (naïve, privileged teenager) belief that knowledge should be free and easily accessible to ambitious high schoolers, but it is paramount to respect intellectual property and someone else's work. I ask that you use this material to its fullest potential, however you like, provided it doesn't completely butcher the philosophy with which I embarked on this project.

<https://www.dropbox.com/s/z2lur71042pjaet/guide.pdf?dl=0>

https://github.com/alwayswimmin/cs_guide

For Rachel

Acknowledgements

I am truly thankful for all my school teachers and USACO instructors, who gave me the knowledge which I wish to pass on through this text. I am grateful for Shankar and Ross, Evan, and Dr. Osborne, who separately made me believe such a task was possible. I must credit Valerie, for she was the one who drove me to really dive into and come to love computer science. Alex and Shwetark were extremely helpful with writing this text, as they provided input and helped me find problems to supplement my work. Finally, I'd like to thank all my friends and avid programmers in the TJ community. This is for all of you, and I hope you find it useful in your studies.

Contents

Acknowledgements	v
Preface	xiii
1 Fundamentals	1
1.1 Comparable	1
1.2 Sorts	2
1.2.1 Merge Sort	2
1.2.2 Quicksort	2
1.3 Binary Search	2
2 Java Standard Library Data Structures	3
2.1 Generics	3
2.2 List	3
2.2.1 ArrayList	4
2.2.2 LinkedList	5
2.2.3 Stack	7
2.2.4 Queue	7
2.3 PriorityQueue	8
2.4 Set	12
2.4.1 TreeSet	12
2.4.2 HashSet	15
2.5 Map	17
2.6 BigInteger	17
2.7 C++ Analogs	17

3	Big Ideas	19
3.1	Search Techniques	19
3.1.1	Complete Search	19
3.1.2	Greedy Algorithm	20
3.1.3	Binary Searching on the Answer	21
3.2	Dynamic Programming	21
3.2.1	Dynamic Programming over Subsets	23
4	Graph Algorithms	25
4.1	Connected Components	25
4.1.1	Flood Fill	26
4.1.2	Union-Find (Disjoint Set Union)	26
4.1.3	Tarjan	28
4.2	Shortest Path	28
4.2.1	Dijkstra	28
4.2.2	Floyd-Warshall	32
4.2.3	Bellman-Ford	32
4.3	Minimum Spanning Tree	33
4.3.1	Prim	33
4.3.2	Kruskal	34
4.4	Eulerian Tour	34
4.5	Maximum Flow	35
4.5.1	Ford-Fulkerson	36
4.5.2	Max-Flow Min-Cut Theorem	37
4.5.3	Edmonds-Karp	38
4.5.4	Dinic	40
4.5.5	Push-Relabel	40
4.5.6	Other Problems with Flow Solutions	44
5	Computational Geometry	45
5.1	Basic Tools	45
5.2	Formulas	45
5.2.1	Area	45
5.2.2	Distance	45
5.2.3	Configuration	45
5.2.4	Intersection	45
5.3	Convex Hull	45

6	Strings	47
6.1	String Hashing	47
6.2	Rabin-Karp	47
6.3	Knuth-Morris-Pratt	47
6.4	Trie	47
6.5	Suffix Data Structures	47
6.5.1	Suffix Tree	47
6.5.2	Suffix Array	47
6.6	Aho-Corasick	47
7	More Stuff	49
7.1	\sqrt{n} Bucketing	49
7.2	Segment Tree	50
7.2.1	Fenwick Tree	53
7.2.2	Lazy Propagation	57
7.2.3	Higher Dimensions	60
7.2.4	Persistent Segment Tree	60
7.3	Queue with Minimum Query	60
7.4	Balanced Binary Search Tree	61
7.4.1	Tree Rotation	61
7.4.2	Red-Black Tree	62
7.4.3	Splay Tree	67
7.4.4	Treap	70
7.5	Euler Tour Technique	70
7.5.1	Euler Tour Tree	71
7.6	2^n Jump Pointers	72
7.7	Fractional Cascading	72
7.8	Heavy-Light Decomposition	72
7.9	Link-Cut Tree	73
7.10	Top Tree	73
7.11	Link-Cut Cactus	73

8	Math	75
8.1	Number Theory	75
8.2	Combinatorial Games	75
8.3	Karatsuba	75
8.4	Matrices	75
8.5	Fast Fourier Transform	75
9	Problems	77
9.1	Bronze	77
9.2	Silver	77
9.2.1	Complete Search	77
9.2.2	Greedy	77
9.2.3	Standard Dynamic Programming	77
9.2.4	Standard Graph Theory	78
9.2.5	Easy Computational Geometry	81
9.3	Gold	81
9.3.1	2^n Dynamic Programming	81
9.3.2	Binary Search	81
9.3.3	Segment Tree	83
9.3.4	More Standard Graph Theory	83
9.3.5	Standard Computational Geometry	84
9.3.6	Less Standard Problems	84
9.4	Beyond	84
9.4.1	Data Structure Nonsense	84

List of Algorithms

1	Union-Find	28
2	Dijkstra	29
3	Floyd-Warshall	32
4	Bellman-Ford	33
5	Prim	34
6	Kruskal	34
7	Eulerian Tour	35
8	Ford-Fulkerson	36
9	Edmonds-Karp	39
10	Push-Relabel (Generic)	43

Preface

You might have heard of Evan Chen’s Napkin, a resource for olympiad math people that serves as a jumping point into higher mathematics.¹ The Wikipedia articles on higher mathematics are just so dense in vocabulary and deter many smart young students from learning them before they are formally taught in a course in college. Evan’s Napkin aims to provide that background necessary to leap right in.

I feel the same way about computer science. For most, the ease of the AP Computer Science test means that the AP coursework is often inadequate in teaching the simplest data structures, algorithms, and big ideas necessary to approach even silver USACO problems. On the other hand, even the best reference books, like Sedgewick, are too dense and unapproachable for someone who just wants to sit down and learn something interesting.² The road, for many, stalls here until college. Everyone should be able to learn the simplest data structures in Java or C++ standard libraries, and someone with problem-solving experience can easily jump right into understanding algorithms and more advanced data structures.

A few important notes, before we begin.

- I’m assuming some fluency in C-style syntax. If this is your first time seeing code, please look somewhere else for now.
- It is essential that you understand the motivations and the complexities behind everything we cover. I feel that this is not stressed at all in AP Computer Science and lost under the heavy details of rigorous published works. I’m avoiding what I call the heavy details because they don’t focus on the math behind the computer science and lose the bigger picture. My goal is for every mathematician or programmer, after working through this, to be able to code short scripts to solve problems. Once you understand how things work, you can then move on to those details which are necessary for building larger projects. The heavy details become meaningless as languages develop or become phased out. The math and ideas behind the data structures and algorithms will last a lifetime.
- It is recommended actually code up each data structure with its most important functions or algorithm as you learn them. I truly believe the only way to build a solid foundation is to code. Do not become reliant on using the standard library (`java.util`, for instance) without understanding how the tool you are using works.

¹In fact, I’m using Evan’s template right now. Thanks Evan!

²Sedgewick, notably, is getting better. Check out his online companion to *Algorithms, 4th Edition*.

Chapter 1

Fundamentals

1.1 Comparable

The interface `Comparable` is how Java can impose an ordering on any object. Implementing `Comparable` requires support to the method `int compareTo(Object o)`. If `a.compareTo(b) < 0`, then `a` should come before `b` in the ordering. If `a.compareTo(b) > 0`, then `a` should come after `b` in the ordering. If `a.compareTo(b) == 0`, then `a` is considered to be equal to `b`. `a.compareTo(b) == 0` should be equivalent to `a.equals(b)`. For example, consider the following class `MyPair`:

```
1 class MyPair implements Comparable {
2     int x;
3     int y;
4     // sort by x-coordinate first
5     public int compareTo(Object o) {
6         MyPair c = (MyPair) o;
7         if(x < c.x) return -1;
8         if(x > c.x) return 1;
9         // if x-coordinates equal, compare y
10        if(y < c.x) return -1;
11        if(y > c.x) return 1;
12        return 0; // equal
13    }
14 }
```

We can also clean this up a bit using generics:

```
1 class MyPair implements Comparable<MyPair> {
2     int x;
3     int y;
4     public int compareTo(MyPair c) {
5         if(x < c.x) return -1;
6         if(x > c.x) return 1;
7         if(y < c.x) return -1;
8         if(y > c.x) return 1;
9         return 0;
10    }
11 }
```

Note that no casting is required anymore.

1.2 Sorts

1.2.1 Merge Sort

The idea behind merge sort is simple. If we are given two equally sized sorted stacks, we could easily combine them into one sorted stack by comparing the top of the two stacks only, since the smallest element remaining must be one of those two.

From here, we divide and conquer. Simply cut the array in half, recurse on each half, and merge the two back together. This runs in $O(n \log(n))$ time.

1.2.2 Quicksort

Quicksort also uses a divide and conquer strategy to achieve an $O(n \log(n))$ amortized bound. We take a random element from the array, move anything less than it to the left of the array, anything right of it to the right of the array, and recurse.

This was previously implemented by `Arrays.sort()` and `Collections.sort()`, but Java now uses either dual-pivot quicksort or timsort.

1.3 Binary Search

Before we begin studying data structures, it is necessary to first understand general search techniques. Given a list, it is straightforward to simply check every element in the list in a linear search.

One way we can make this faster is by imposing an ordering on the list. Then inserting elements becomes slower because we can no longer just pop the element to the the end of the list. However, if we are asked to find an element, we know roughly where in the list to search because we can compare the element to other elements in the list.

Consider an array of sorted elements. The array supports $O(1)$ access to any element in the array. We'll take a look at the middle element in the array. Depending on whether or not the element to be searched is larger, smaller, or equal to the middle element, we can guarantee we can eliminate half of the array. Then finding an element is an $O(\log n)$ operation on a sorted array. This can be done either iteratively or recursively.

Chapter 2

Java Standard Library Data Structures

The purpose of this chapter is to provide an overview on how the most basic and useful data structures work. The implementations of most higher-level languages already coded these for us, but it is important to know how each data structure works rather than blindly use the standard library.

More technical explanations of all of these can be found, mostly under the package `java.util`, in the Java API.

I strongly believe that Java is better than C++ for beginning programmers. It forces people into good coding habits, and though the lack of pointers initially frustrated me, it really does make learning general concepts like `LinkedLists` much easier, as the intricacies of the C++ pointer no longer distract from the larger idea.

2.1 Generics

You may or may not have seen these used before:

```
ArrayList<Integer> al = new ArrayList<Integer>();
```

This means that `al` is an `ArrayList` of `Integers`. We can only add `Integers` into the `ArrayList`, and anything removed from the `ArrayList` is guaranteed to be an `Integer`. I don't think the beginning programmer needs to know how to necessarily code a class that supports generics. However, you are going to use the standard library, so you should know how to use a class that does support generics. I think standard classes are relatively straightforward to use but can be annoying to actually implement.

For simplicity, all data structures in this chapter will store Strings.

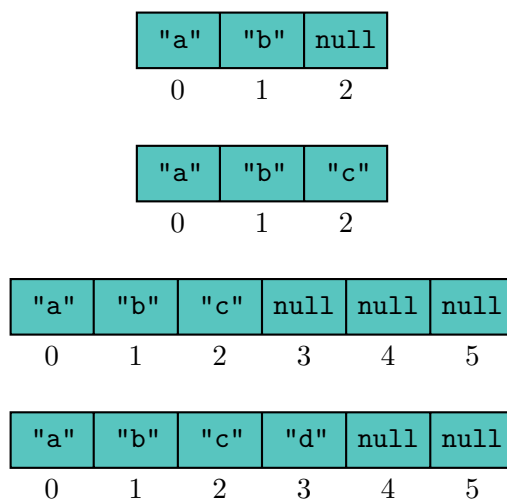
2.2 List

A `List` is a collection of objects with an ordering. Users of a list have control over where in the list each object is and can access a specific element by its index, like in an array.

2.2.1 ArrayList

What is nice about an array? We can access or change any element we want in the array in $O(1)$ time. The problem is that an array has fixed length. It's not easy to append an element to the end of an array.

The fix to this is pretty simple. Why not just make a bigger array, and copy everything over to the new array? Then there's more room at the end to add a new element. This is the idea behind an ArrayList. If the backbone array runs out of space, we create a new array with double the size and keep going as if nothing happened.



We see that there is still room in the array to add "c", but to add more elements to the list, we must use a new array with double the length.

It's important to note that any given insertion to the ArrayList is either $O(n)$ or $O(1)$, but there is only one $O(n)$ insertion for every $O(n)$ $O(1)$ insertions, so we still average out to constant time.

- `boolean add(String s)` – add an element to the end of the list. (By convention, this returns `true` if the addition was successful, and `false` otherwise.)
- `void add(int i, String s)` – shift everything from position `i` onward down by one, and add `s` at position `i`.
- `boolean contains(Object o)` – return `true` if `o` is stored in the ArrayList, and `false` otherwise. Remember to cast `o` to a `String`.
- `String get(int i)` – return the element stored at index `i`.
- `boolean isEmpty()` – return `true` if `size() == 0`.
- `String remove(int i)` – remove and return the element at index `i` from the list. (Why is this annoying?)
- `String set(int i, String s)` – replace the element stored at index `i` with `s`, and return the element that was originally at position `i`.

- `int size()` – note that this is not just the length of the backbone array.

The `get` and `set` functions are very nice. They are easy to code and run in constant time. These are the bread and butter of any array. Adding at the end of the `ArrayList` is nice as well. `contains` is a pain, as it is $O(n)$, and adding to and removing from early in the list are more annoying.

If this is your first time seeing an `ArrayList`, I would suggest coding up your own. It should be relatively straightforward.

2.2.2 LinkedList

Arrays are nice for accessing, say, the seventh element in the list. We extend this to an `ArrayList` to implement adding and removing elements to and from the end of the list nicely. Removing elements from the beginning of the list, however, is cumbersome.

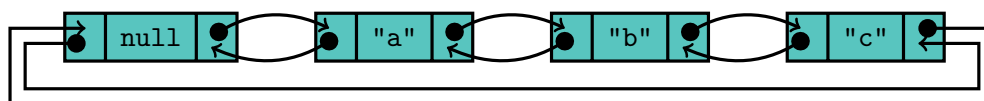
The `LinkedList` attempts to remedy this. It trades $O(1)$ access to any element in the list for an easier way to remove elements from either end of the list easily. Consider a chain of paper clips:



1

It's easy to add or remove more paper clips from either end of the chain, and from any given paper clip, it's easy to access the paper clip directly previous or next to it in the chain. If we needed the seventh paper clip in the chain, we'd need to manually count, an $O(n)$ operation. However, if we then needed to remove that paper clip from the chain, it wouldn't be that hard, assuming we kept a finger on the seventh paper clip.

The standard library implements a cyclical doubly-linked list, with a dummy head. It looks something like this:



¹http://img.thrfun.com/img/078/156/paper_clip_chain_s1.jpg

We see that each node maintains a pointer to its next neighbor and its previous neighbor, in addition to containing the String it stores. We can store this data in a class like the following:

```
1 class ListNode {  
2     ListNode prev, next;  
3     String s;  
4 }
```

If we were to insert an element after a `ListNode a`, it is necessary to update all pointers:

```
1 ListNode b = new ListNode();  
2 b.prev = a;  
3 b.next = a.next;  
4 b.next.prev = b;  
5 a.next = b;
```

Since the `LinkedList` is symmetric, inserting an element before a node is also easy. To add something to the end of the list, simply add it before the dummy head. From here it should not be too hard to implement all the important functions of a `LinkedList`.

- `boolean add(String s)` – add to the end.
- `void add(int i, String s)` – this is linear. Don't forget about 0-indexing.
- `void addFirst()`
- `boolean contains(Object o)`
- `String get(int i)`
- `String getFirst()`
- `String getLast()`
- `boolean isEmpty()`
- `String remove()` – remove the last element.
- `String remove(int i)`
- `String removeFirst()`
- `String set(int i, String s)`
- `int size()`

With a `LinkedList` implemented, two other data structures immediately follow.

2.2.3 Stack

A stack is literally a stack. If we have a stack of papers, we can *push* things on the top and *pop* things off the top. Sometimes we *peek* at what's on the top but don't actually remove anything. We never do anything with what's on the bottom. This is called *LIFO*: Last In, First Out.

- `String push(String s)` – pushes the item on the top of the stack, and returns the same element.
- `String pop()` – removes the item at the top of the stack and returns it.
- `String peek()` – returns the top element but does not remove it.

Java implements a Stack using an ArrayList-like structure. This works just as well, and is faster in practice, but I prefer the LinkedList structure as a mathematical concept as it is more elegant and more easily customizable.

2.2.4 Queue

A queue is like a lunch line. We *add* things to the end and *poll* things from the front. Sometimes we *peek* at the front but don't actually remove anything. The first person in line gets served first. This is called *FIFO*: First In, First Out.

- `boolean add(String s)`
- `String poll()` – removes the item at the front of the queue and returns it. Same thing as `remove()`.
- `String peek()`

In Java, `Queue` is an interface. This means that we cannot instantiate a `Queue`, so the following statement is illegal.

```
1 Queue<String> q = new Queue<String>();
```

Instead, we must do something like this:

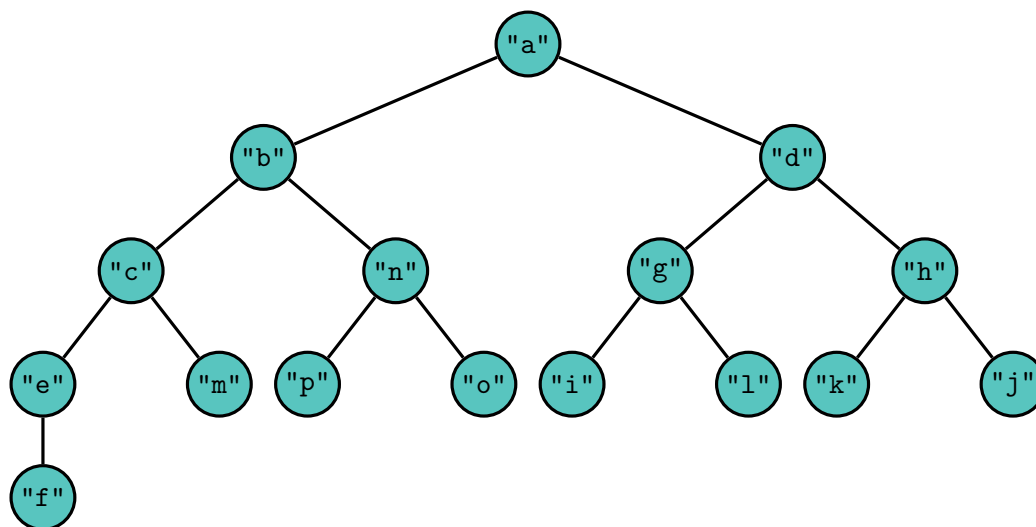
```
1 Queue<String> q = new Queue<LinkedList>();
```

This is legal because `LinkedList` implements `Queue`. `Queue`, however, does not extend `List`, so it is somewhat untruthful to place `Queue` under `List` in this book. However, I do want to stress that the `LinkedList` is the standard FIFO queue.

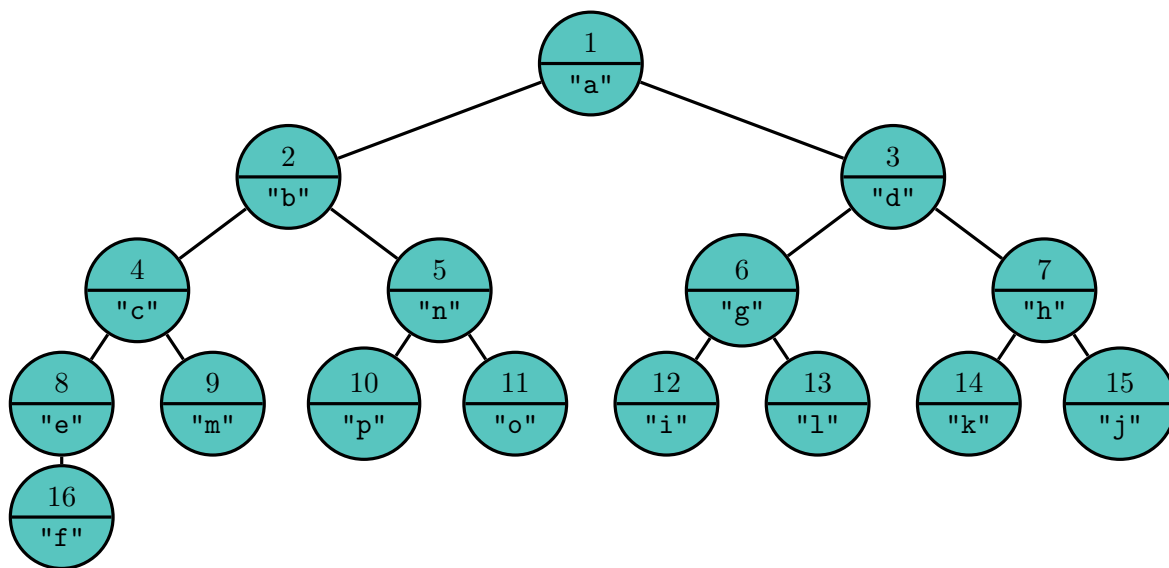
2.3 PriorityQueue

Quite often a FIFO queue is not always desirable. Maybe the String I want to remove at every given point is the one that is lexicographically least. a PriorityQueue is a Queue that allows us to do this using what is known as a *heap*.

A min heap is a tree such that every node is smaller than or equal to all of its children. Pictured is a complete binary min heap, which will be of most use to us.



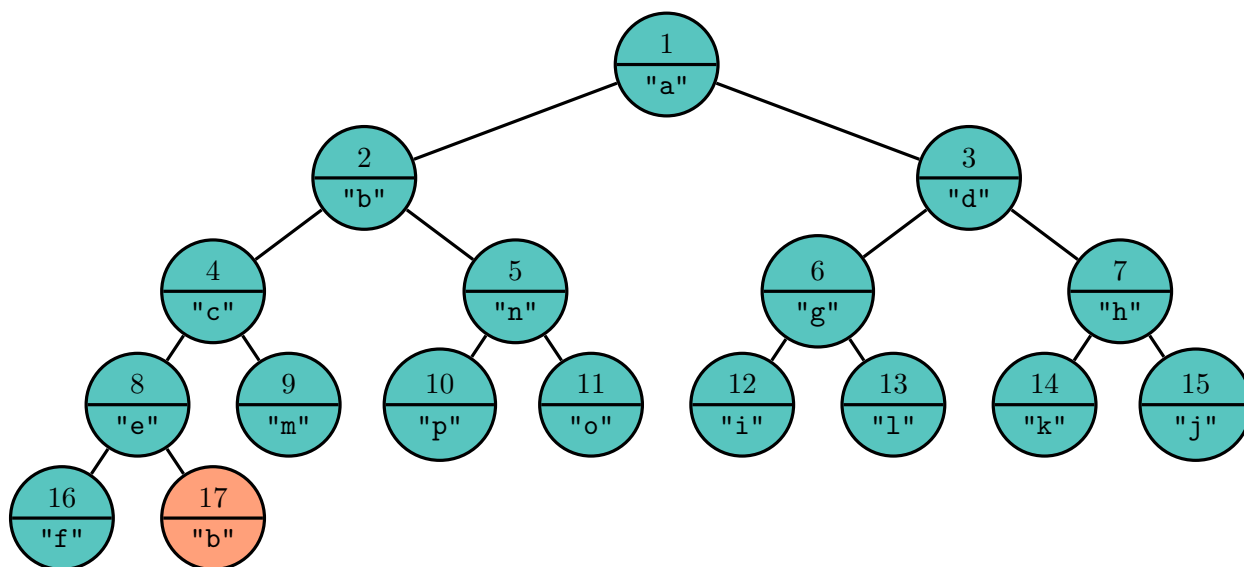
We see that the root of the tree will always be the top element. It is tempting to use a container class with a pointer to its left and its right child. However, we have a much nicer way to store *complete* binary trees with an array. Consider the following numbering of the nodes:



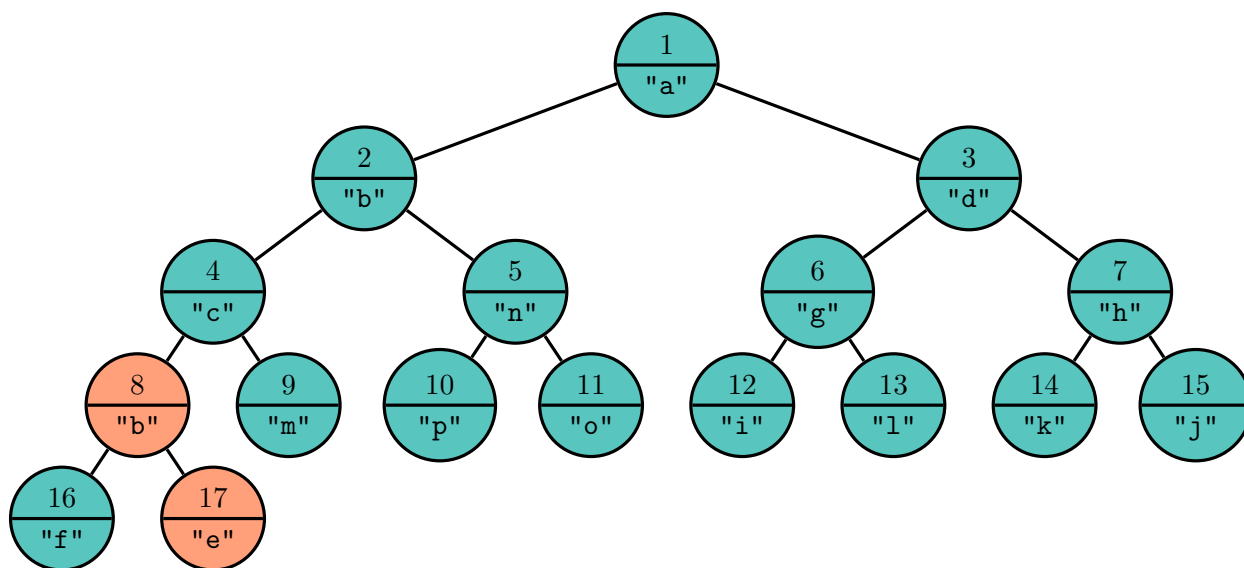
We see that every number from 1 to 16 is used, and for every node, if the index associated with it is i , the left child is $2i$, and the right child is $2i + 1$. This leads to a very natural implementation of the tree in an array:

null	"a"	"b"	"d"	"c"	"n"	"g"	"h"	"e"	"m"	"p"	"o"	"i"	"l"	"k"	"j"	"f"
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

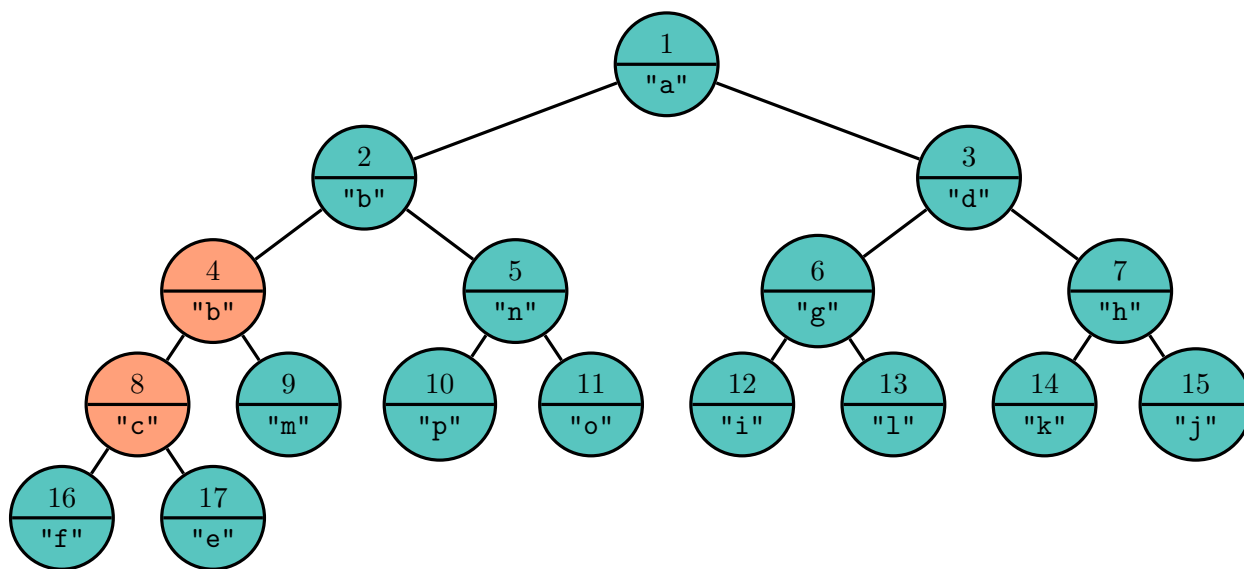
How do we add elements to our heap, while maintaining the heap qualities? Well, let's just add it to the very end and see what we get. Suppose we are to add "b" to the tree.



Well, "b" comes before "e" in the alphabet, so let's swap the nodes. We are guaranteed that "b" should come before the other child (in this case, "f") by the transitive property.

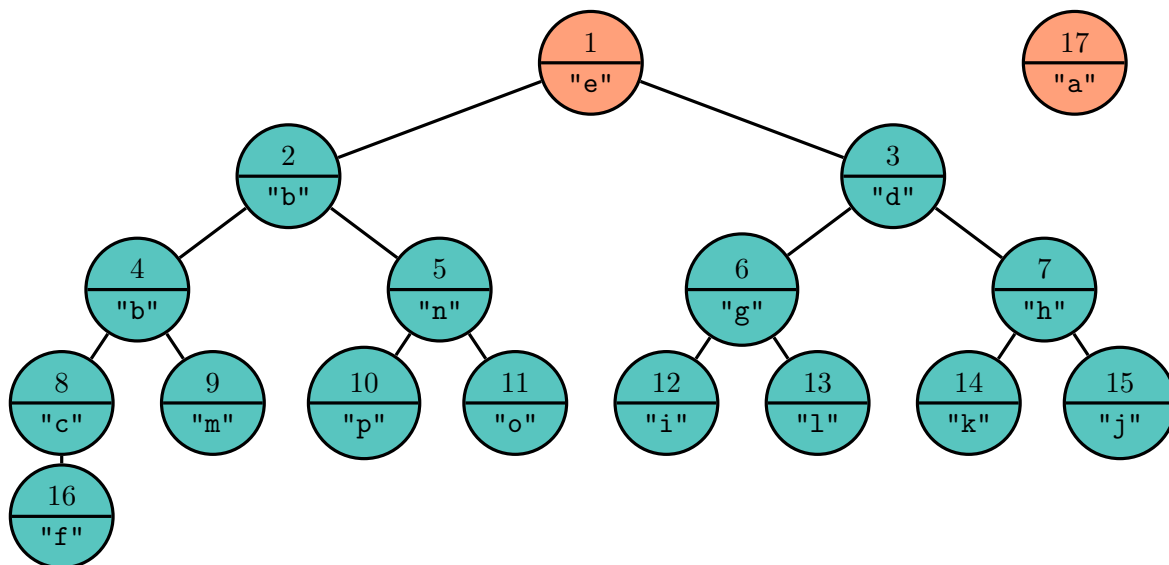


One more swap...

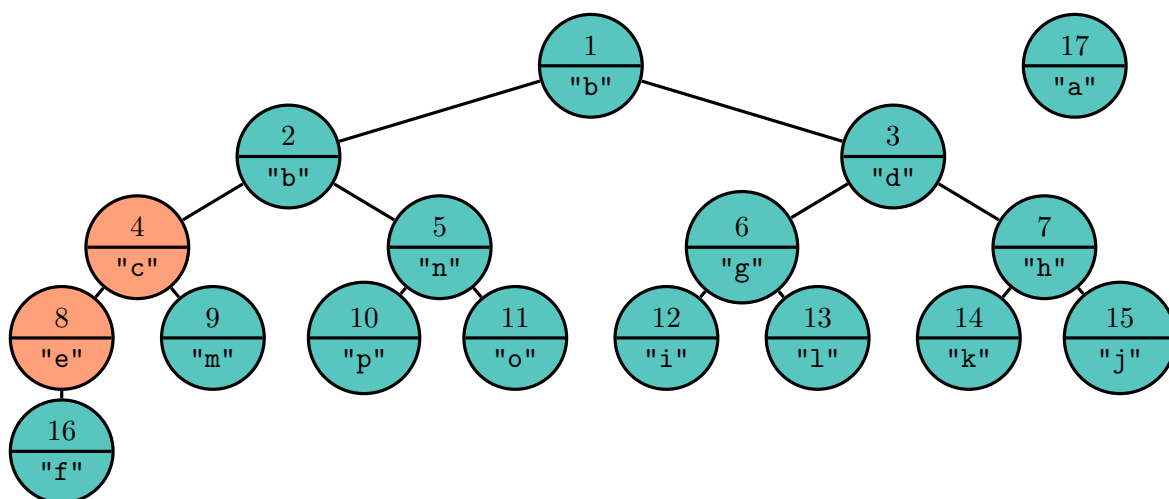
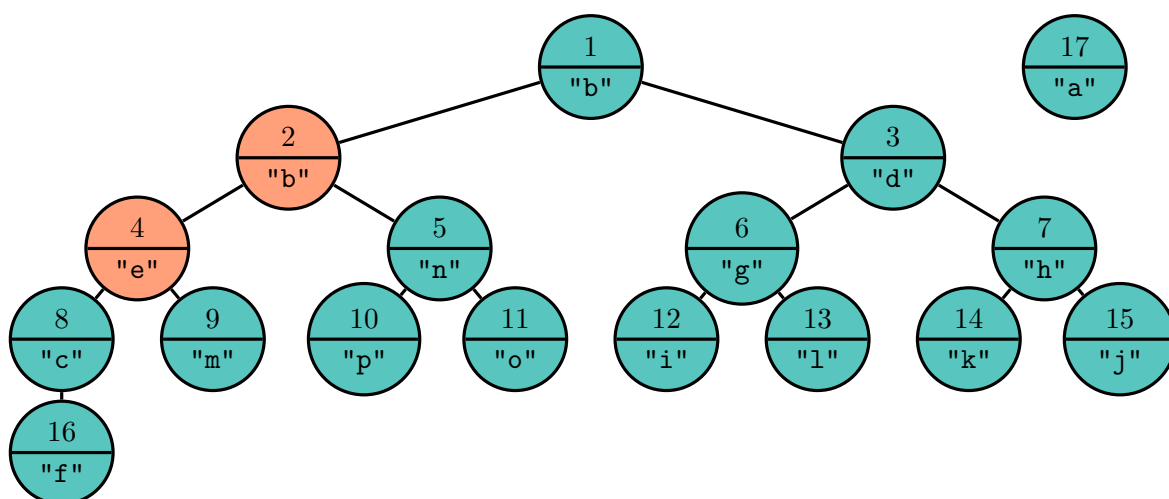
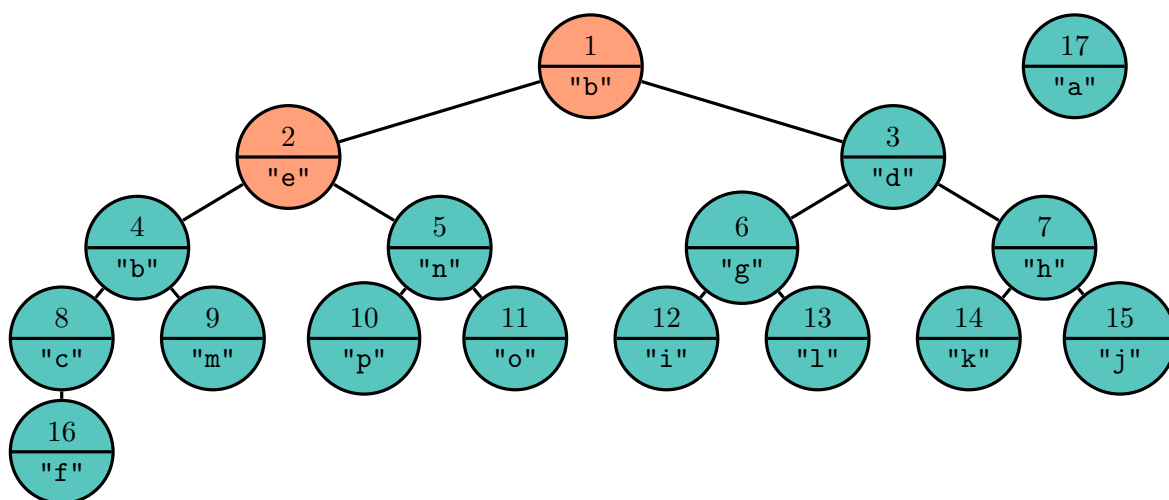


And now we have the heap property restored. As the tree has depth at most $\log n$, this process is $O(\log n)$.

To remove the root from the heap, we replace the root with the last leaf:



We perform a series of swaps to restore the heap property. We always want to choose the smaller child to swap until the heap property is satisfied.



And we are done. Once again, this takes at most $\log(N)$ swaps. This idea can be extended to removing or changing the value of any node we'd like from a tree – this is particularly useful for Dijkstra later.

Remember to implement your heap in an array or ArrayList!

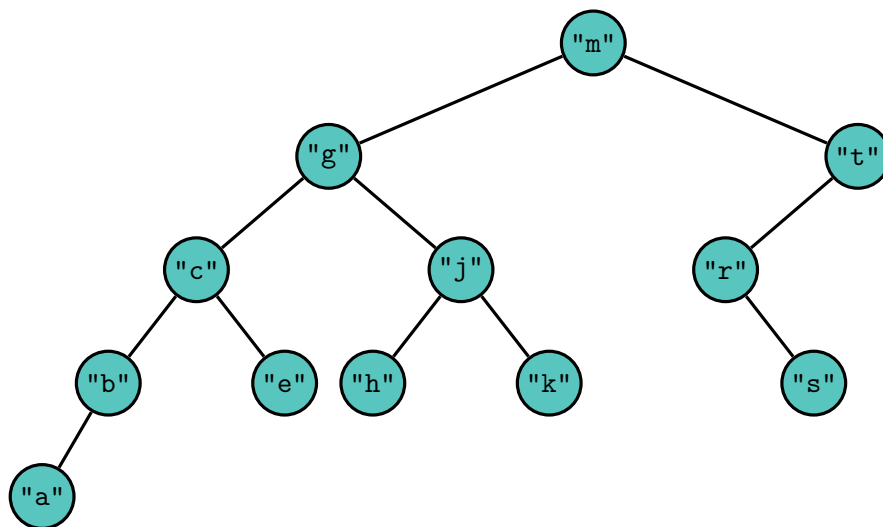
- `boolean add(String s)`
- `boolean isEmpty()`
- `String poll()`
- `String peek()`
- `int size()`

2.4 Set

A Set is a collection of objects with no duplicate elements. Set is a Java interface. Note that the data structures discussed in this section can be extended to become multisets, but Java implementations of these explicitly disallow multiplicity.

2.4.1 TreeSet

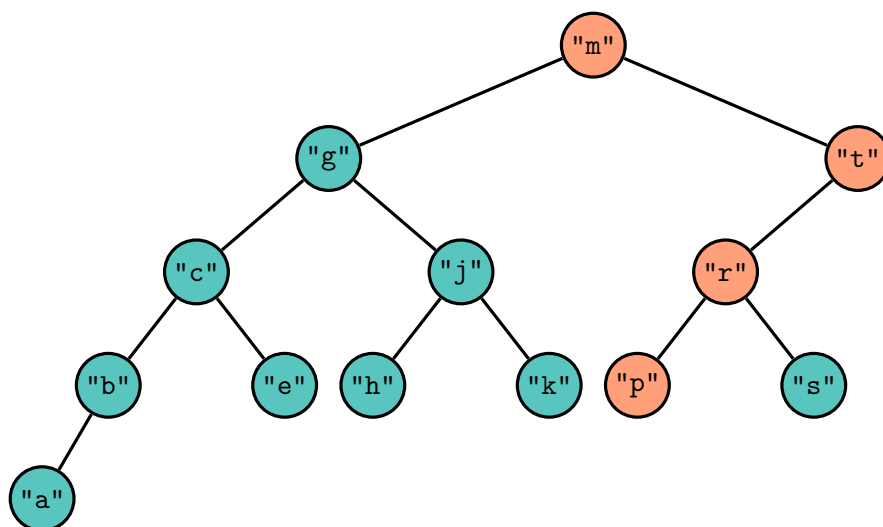
A TreeSet is Java's implementation of a *binary search tree* (BST). A binary search tree is a tree where every node is greater than every node in its left subtree and less than every node in its right subtree. As with a heap, to use a BST, we need to impose some kind of ordering on the elements stored.



The tree need not be complete. Because it is not complete, there is no way to nicely bound the size of the array we would need if we were to use the same storage method as with the heap. Thus, we are forced to use a `TreeNode`, with left and right pointers. This is also problematic when determining guarantees on time complexities later, but the ways to solve this problem are pretty complicated so we'll ignore them for now.

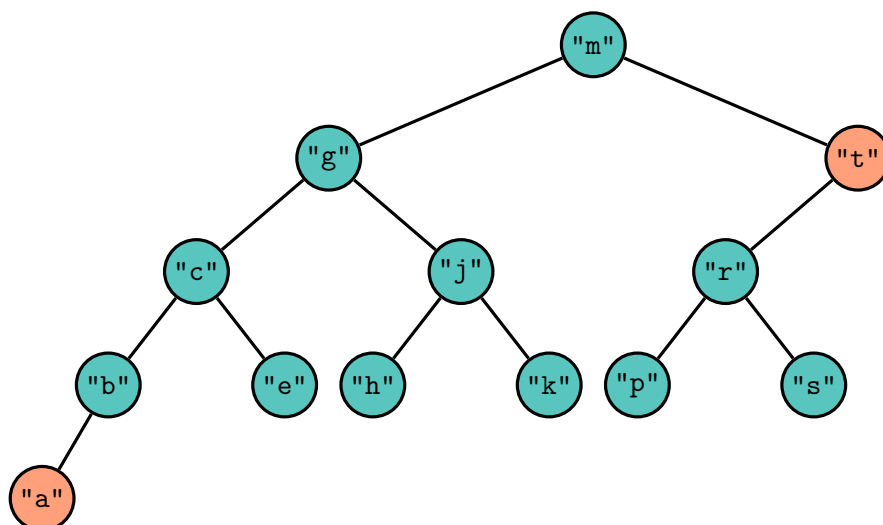
Given the name of the tree, searching for an element within the tree is quite natural, and similar to a binary search. Compare the element to be searched for with the current node. If they are equal, we are done; otherwise, search the appropriate left or right subtree. As with most structures and algorithms with a binary search structure, this operation lends itself nicely to recursion. If the tree is reasonably nice, we expect to complete this in $O(\log n)$ time, but searching can be as bad as linear if the tree looks like a LinkedList.

Adding an element is also natural. As our tree represents a set, it will not contain the same element twice. We trace down until we hit a null pointer, and add the element in the appropriate spot. Let's add a "p" to the BST:

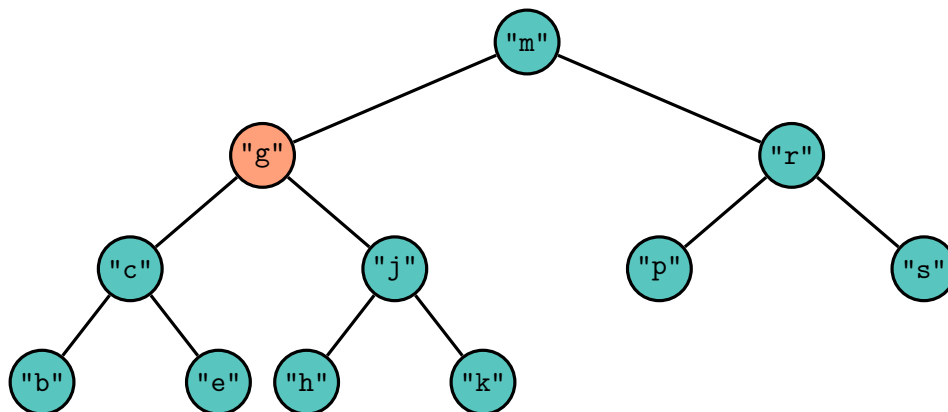


Deleting an element is the annoying part. Unfortunately, there's not much we can do besides casework.

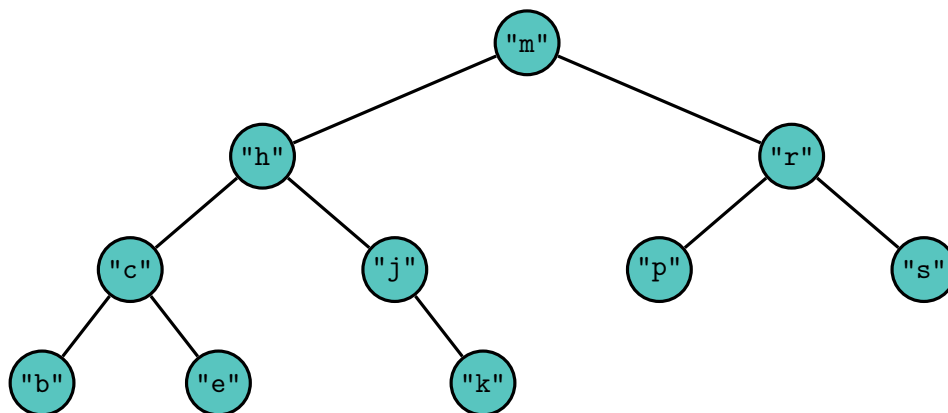
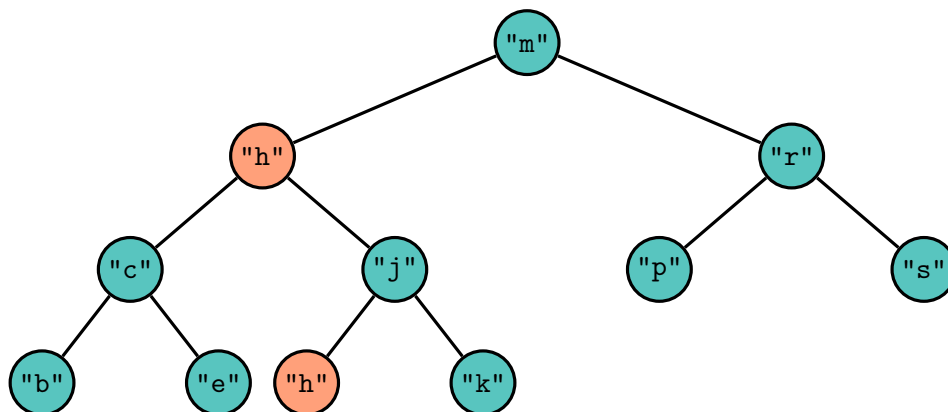
Removing a leaf, like "a", from the tree is very easy. Removing a node with only once child, like "t", is also relatively straightforward.



Now, removing an element with two children is tricky. We'll try to remove "g". Consider the least element in the right subtree of "g", which in this case is "h". We find "h" by always choosing the left child on the right subtree until we cannot go any further. This must be the least element.



Note that "h" has either no children or only one child, and that nodes like these are easy to remove. We then change the value of the node containing "g" to "h", which is legal since "h" is the least element, and remove "h" from the right subtree, and we are done.



A standard BST has $O(\log n)$ operations if the tree is “nice”, but each operation can be $O(n)$ in the worst case. We need to find a way to automatically balance the BST such that we avoid linear time complexities.

A red-black tree is a self-balancing BST that guarantees $O(\log n)$ operations by making sure the height of the tree grows logarithmically. It is implemented in Java’s `TreeSet`, so while the BST I described above does not guarantee nice time bounds, Java’s implementation does.

I don’t think learning exactly how a red-black tree works is particularly useful for the beginning programmer. How exactly a red-black tree works, together with some more balanced binary search trees which are useful on the competitive scene, are covered in a later chapter.

Here are some notable functions `TreeSet` implements. You don’t need to implement everything – `add()`, `remove()`, and `contains()` are the most important.

- `boolean add(String s)`
- `String ceiling(String s)` – the least element in the set greater than or equal to `s`, or `null` if there is no such element.
- `boolean contains(Object o)`
- `String first()` – the least element in the set.
- `String floor(String s)` – the greatest element in the set less than or equal to `s`, or `null` if there is no such element.
- `String higher(String s)` – the least element in the set strictly greater than `s`, or `null` if there is no such element.
- `boolean isEmpty()`
- `String last()` – the greatest element in the set.
- `String lower(String s)` – the greatest element in the set strictly less than `s`, or `null` if there is no such element.
- `boolean remove(Object o)`
- `int size()`

Since the `TreeSet` is ordered, iterating over a `TreeSet` will also be in order.

2.4.2 HashSet

A `HashSet` is a way for us to store objects when we do not require a natural ordering on the set. As with the `TreeSet`, we want to be able to check whether an element is in our set or not quickly. We do this with the help of a *hash function*. Every Java Object supports the `hashCode()` function. We usually want to map an object with an integer hash, so that we can store the values in an array. For example, let us define the following hash for Strings:

```

1 public int hashCode() {
2     int hash = 0;
3     for(int k = 0; k < length(); k++) {
4         hash *= 31;
5         hash += (int) (charAt(k));
6     }
7     return hash;
8 }

```

`a.equals(b)` should imply `a.hashCode() == b.hashCode()`. This function produces the same result as the actual `hashCode()` function in the `String` class. However, this is not quite what we want for our `HashSet`, because in the end we wish to be able to store the objects in some kind of array. This hash not only returns integers that can be very large, they can also be negative, and thus are not suitable as array indices. The natural way to fix this is to take the hash modulo the size of the array we want.

```

1 String[] table = new String[10007];
2 int index(String s) {
3     int i = s.hashCode() % table.length;
4     if(i >= 0)
5         return i;
6     return i + table.length;
7 }

```

We chose the number 10007 because it is a prime number, and primes are generally nice when taking a number modulo something else, as integers modulo a prime form a field. Remember that a negative number % another number is not necessarily positive, so we need to be a little careful.

From here, adding an element to the `HashSet` and checking if an element is contained both seem straightforward:

```

1 boolean add (String s) {
2     table[index(s)] = s;
3     return true;
4 }
5 boolean contains(Object o) {
6     int i = index((String) o);
7     return table[i] != null && table[i].equals(o);
8 }

```

`null` is always annoying to deal with, and will have to be handled separately.

However, one problem quickly arises. Two `Strings` may map to the same index in the array. We call this a *collision*. The easiest way to handle a collision is by *chaining*. We change the hash table to store a `LinkedList` instead of a single element in the event of a collision. Java once implemented this method of resolving collisions, but recently changed the `LinkedList` to a `BST` in Java 8.

- `boolean add(String s)`
- `boolean contains(Object o)`
- `boolean isEmpty()`

- `boolean remove(Object o)`
- `int size()`

2.5 Map

A Map is simply an extension of a Set. It stores a mapping that takes a key to a value. Map is a Java Interface. Generics for Maps therefore have two arguments. Consider the following Map from Strings to Strings.

```
1 Map<String, String> email = new TreeMap<String, String>();  
2 email.put("Samuel Hsiang", "samuel.c.hsiang@gmail.com");
```

The keys of a Map form a Set, though the Values need not be unique.

The TreeMap is the Map variant of the TreeSet; similarly, the HashMap is the Map variant of the HashSet.

All useful Set functions have a Map counterpart. The following additional functions are of use.

- `String get(Object k)` – return the value assigned to `k`.
- `Set<String> keySet()` – returns the set of keys.
- `Integer put(String k, String v)` – assigns the value `k` to the key `v`, and returns the old value assigned to `k`.

2.6 BigInteger

BigInteger is in `java.math` for times when `int` and `long` just aren't large enough. The way BigInteger works is it stores a number as an array of ints. Each value in the array represents a digit in some very large base. Addition and subtraction can be done in the standard way. Generally multiplying two BigIntegers is not necessary on contests, but it can be sped up using Karatsuba or the FFT.

2.7 C++ Analogs

- `ArrayList` – vector
- `LinkedList` – list or deque
- `Stack` – stack
- `Queue` – queue
- `PriorityQueue` – `priority_queue`, but note that `priority_queue` pops *max* element first
- `TreeSet` – set

- `HashSet` – `unordered_set`, C++11
- `TreeMap` – `map`
- `HashMap` – `unordered_map`, C++11
- `BigInteger` – no equivalent

Chapter 3

Big Ideas

In this chapter we'll discuss some general ideas for solving problems. Starting in this chapter I'm going to shift from language-specific terms, like `PriorityQueue` and `TreeMap`, to more general terms, like binary heap and binary search tree. Algorithms I present will no longer be in the form of concrete Java code but rather in a more abstract pseudocode.

3.1 Search Techniques

There are many different general strategies for finding the answer to a problem.

3.1.1 Complete Search

Sometimes the best way is simply to try everything. This could be the intended solution (check the complexity of a complete search and compare it against the time limits), or we may be just coding a brute force to squeeze out a few extra points from an intractable problem at the end of a contest. Either way, the order in which we search can make or break the code.

Suppose we wanted to solve the game of chess. The placement of the pieces on the board and a toggle for whether black or white is to move represents a game state. The set of legal moves maps this game state to one set of states and another set of states to this game state.

We can therefore think of the different states as vertices in a very large directed graph. Methods we use for solving chess, or any other problem, are identical to ways of traversing graphs, especially trees.

Depth-First Search

Depth-first search is the most simple of the searches. A depth-first search called on a vertex v in the search tree recursively calls itself on each of the children of v . To go back to the chess example, a DFS would begin with one possible first move, like a3, and test *every possible move sequence* beginning with a3 before moving on to a second possible first move, like a4.

A quick note on recursive processes: The *run-time stack* keeps track of our position and data that go out of scope when we jump into a new function. This stack allows us to call functions within

functions safely. Recursion is dangerous in contest programming because the run-time stack is slow and can easily balloon in size, crashing the program. For this reason do not hesitate to use your own stack in an iterative process in practice, though I often describe processes as recursive as they are easier to understand, code, and debug in this form.

Now it is painfully obvious that strictly using a DFS is not preferable in computer chess. The computer would certainly never finish testing every move sequence beginning with a3 and therefore would not consider more standard, and likely better, moves like e4.

Breadth-First Search

Breadth-first search traverses the search tree level by level. We keep track of a queue that stores each level's state. At each step we pop off the first element in the queue and add the states that it can reach to the end of the queue. In this way, we explore those states only after we explore all the states associated with a lower level. In the chess example, we would store all 20 possible first moves in the queue, and for each first move, we add all possible second moves to the end of the queue.

This means that we search all possible states of a certain depth at roughly the same time, but we need to store the data associated with each of these states, and this can be quite costly in terms of memory.

BFS is the better choice over DFS when we are asked for the “smallest” answer, which is often associated with the lowest level.

Depth-First Search with Iterative Deepening

Depth-first search with iterative deepening is a DFS that only searches up to a certain level in the search tree before stopping and heading back to lower levels. If it turns out the search did not find anything in the first N levels, we broaden the search to the first $N + 1$ levels with another DFS, hence iterative deepening. This method of searching is slower than a BFS, but maintains the nice BFS property of finding the “smallest” answer first while shedding the memory harness that holds the BFS back.

DFSID is the closest to what we do when analyzing an actual chess game. We make a move in our minds and test its performance against whatever the opponent might make as his move, tracing the game around 3-4 moves deep before testing another possible move.

3.1.2 Greedy Algorithm

The main problem with any of the searches described above is they are exponential in nature. A *greedy algorithm* is one that takes the “best” possible option at each step, essentially disregarding any other possible option. To find the best solution for $n = 4$, we consider only the best solution for $n = 3$ and no other possible solution. The approach of maximizing at each step clearly does not always work, as the locally optimal choice is not necessarily globally optimal. Always moving in the direction of a target, for example, fails if there is a wall in the way. However, if the greedy algorithm can solve a problem, the code runs very quickly.

One problem that cannot be solve using the greedy approach is the integer knapsack problem. It's important to be able to catch an incorrect greedy algorithm, and one classic example involves

finding the most efficient way to make change. Given a sequence of coin denominations (d_i , where $1 \leq i \leq n$), with $d_1 = 1$, find the smallest number of coins necessary that sum to some value V . The greedy approach takes the most valuable coin that doesn't overshoot V and adds it to our set until we achieve V . One counterexample is $v_1 = 1$, $v_2 = 3$, $v_3 = 4$ and $V = 6$. The greedy algorithm chooses $4 + 1 + 1$, which is worse than $3 + 3$. Note however that greedy algorithm still always works for some sets of coin values, like United States currency. Don't let the fact that the greedy algorithm works for the most obvious example fool you.

The greedy algorithm combined with ideas from dynamic programming constitute *best-first search*, a category of searches that includes the Dijkstra algorithm for shortest paths.

3.1.3 Binary Searching on the Answer

Suppose that the problem we need to solve is finding the minimum number M such that some property holds. That is, the property holds for any $x \geq M$ but does not hold for $x < M$. Perhaps the best approach we have so far for finding this M is simply trying all the numbers from 1 to M . However, this linear search is clearly inefficient.

The nature of this problem should remind you of some other search technique. Oftentimes, with problems of this property, it is easy to check whether some condition holds for some given x . In this case, it is much easier to binary search on M rather than find it some other way.

3.2 Dynamic Programming

The idea behind dynamic programming is to avoid doing the same thing twice. Two nodes in the search tree described in the complete search techniques might very well represent the same state. For example, two different sets of initial moves could result in the same chessboard configuration, and if we already calculated who has the winning or losing position in that configuration, we ought to remember that fact somehow so we don't need to calculate it again. Of course, however, actually keeping track of every possible game state is intractable by memory constraints.

Here's a much more reasonable problem. Given a sequence of $N \leq 10,000$ integers, what is the maximum decreasing subsequence? A subsequence does not have to consist of consecutive terms in the original sequence.

6	9	8	4	7	5
1	2	3	4	5	6

The natural complete search approach would be to use recursion, or DFS. When we process an element in the list, we recursively process all elements that come after it and choose the one that gives the maximum subsequence.

```

function PROCESS( $i$ )
   $max \leftarrow 0$ 
  for  $j \equiv i + 1, N$  do
    if  $value(i) > value(j)$  then
       $x \leftarrow$  PROCESS( $j$ )

```

```

    if  $x > \text{max}$  then
         $\text{max} \leftarrow x$ 
    return  $\text{max} + 1$ 

```

However, this algorithm is exponential. In the worst case, it is $O(2^N)$. We notice a lot of repetition: processing the 9 in the list above, for example, requires finding the longest subsequences beginning with 8, 4, 7, and 5, while processing 8 requires finding subsequences for 4, 7, and 5. It seems silly to do the same task twice, so we'll keep track of the length of the longest subsequence in a separate array.

```

function PROCESS( $i$ )
    if  $i$  has already been processed then
        return  $dp(i)$ 
     $\text{max} \leftarrow 0$ 
    for  $j \equiv i + 1, N$  do
        if  $\text{value}(i) > \text{value}(j)$  then
             $x \leftarrow \text{PROCESS}(j)$ 
            if  $x > \text{max}$  then
                 $\text{max} \leftarrow x$ 
     $dp(i) \leftarrow \text{max} + 1$ 
    return  $\text{max} + 1$ 

```

This reduces the complexity of the algorithm to $O(n^2)$. Note that to process an index, we must process first all later indices. This imposes a natural ordering in which to process the indices: in reverse. This idea lends itself to a nice iterative solution.

```

for  $i \equiv N, 1$  do
     $\text{max} \leftarrow 0$ 
    for  $j \equiv i + 1, N$  do
        if  $\text{value}(i) > \text{value}(j)$  then
            if  $dp(j) > \text{max}$  then
                 $\text{max} \leftarrow dp(j)$ 
     $dp(i) \leftarrow \text{max} + 1$ 

```

▷ i goes in reverse

The answer to the original problem is then the maximum value of $dp(i)$ for all i . For this specific problem, it's relatively easy to speed up the algorithm to $O(\log n)$ by replacing the linear search with something else.

The integer knapsack problem is another example where dynamic programming may be useful. *Knapsack problems* are a family of problems with the following form:

We are given a list of K objects each assigned an availability, a size, and a value. We have a total amount of “space” available in our knapsack and need to find the set of objects from our list that maximizes the total the value of objects in the set such that the total size does not exceed the space and the number of times we take one particular object does not exceed its availability.

Dynamic programming yields a straightforward $O(NK)$ solution. See if you can find it. Note, however, if N is very large, this solution is no longer practical. In general, the knapsack problem is NP-complete, so don't think dynamic programming works on everything!

3.2.1 Dynamic Programming over Subsets

Consider the following problem:

(USACO December 2014, guard) Farmer John and his herd are playing frisbee. Bessie throws the frisbee down the field, but it's going straight to Mark the field hand on the other team! Mark has height H ($1 \leq H \leq 1,000,000,000$), but there are N cows on Bessie's team gathered around Mark ($2 \leq N \leq 20$). They can only catch the frisbee if they can stack up to be at least as high as Mark. Each of the N cows has a height, weight, and strength. A cow's strength indicates the maximum amount of total weight of the cows that can be stacked above her.

Given these constraints, Bessie wants to know if it is possible for her team to build a tall enough stack to catch the frisbee, and if so, what is the maximum safety factor of such a stack. The safety factor of a stack is the amount of weight that can be added to the top of the stack without exceeding any cow's strength.

We can try the $O(N!)$ brute force, trying every permutation of cows possible. However, this is far too slow. $N \leq 20$ hints at an exponential solution, so we think of trying every possible subset of the cows. Given a subset S of cows, the height reached is the same, so perhaps we sort the subset by strength, and put the strongest cow on the bottom. We see that this greedy approach fails: suppose that the first cow has weight 1 and strength 3 and the second cow has weight 4 and strength 2. Greedy would tell us to put the first cow on the bottom, but this fails, while putting the second cow on the bottom succeeds.

When greedy fails, the next strategy we look at is dynamic programming. To decide whether S is stable, we have to find whether there exists a cow j in S that can support the weight of all the other cows in S . But how do we know whether the set $S \setminus \{j\}$ is stable? This is where dynamic programming comes in.

This leads to a $O(N2^N)$ solution. This seems like a pain to code iteratively, but there is a nice fact about subsets: there is a cute bijection from the subsets of $\{0, 1, 2, \dots, N-1\}$ to the integers from 0 to $2^N - 1$. That is, the subset $\{0, 2, 5, 7\}$ maps to $2^0 + 2^2 + 2^5 + 2^7 = 165$ in the bijection. We call this technique *masking*. We require all the subsets of S to be processed before S is processed, but that property is also handled by our bijection, since subtracting a power of 2 from a number decreases it. With a little knowledge of bit operators, this can be handled easily.

```

for  $i \leftarrow 0, 2^N - 1$  do                                ▷  $i$  represents the subset  $S$ 
     $dp(i) \leftarrow -1$ 
    for all  $j \in S$  do                                       ▷  $i \& (1 \ll j) \neq 0$ 
         $alt \leftarrow \min(dp(i - 2^j), strength(j) - \sum_{k \in S \setminus \{j\}} weight(k))$ 
        if  $dp(i) < alt$  then
             $dp(i) \leftarrow alt$ 

```

$\&$ is the bitwise and function, while \ll is the left shift operator.

Brian Dean compiled some standard dynamic programming problems with animations and analyses. Practice dynamic programming here: http://people.cs.clemson.edu/~bcdean/dp_practice/

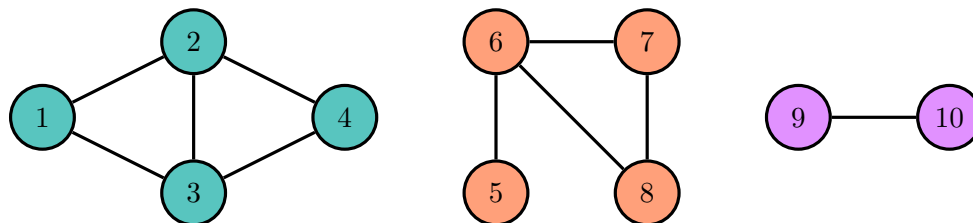
Chapter 4

Graph Algorithms

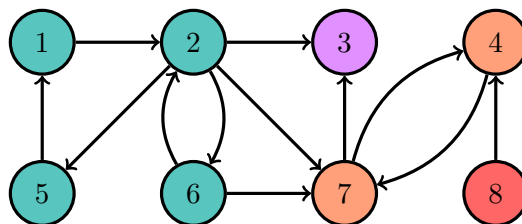
In this chapter we explore some famous graph theory results.

4.1 Connected Components

A *connected component* of an undirected graph is a subgraph such that, for any two vertices in the component, there exists a path from one to the other. The diagram illustrates three connected components of a graph, where each vertex is colored together with its associated component.



A *strongly connected component* of a directed graph is a subgraph such that every vertex in the component can be reached from any other vertex in the component.



Finding the connected components of an undirected graph is a straightforward problem, while finding the strongly connected components of a directed graph is more complicated.

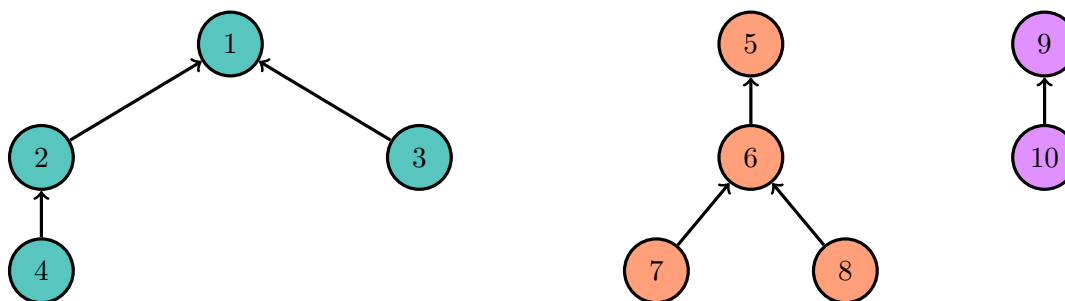
4.1.1 Flood Fill

Really any kind of search method solves the undirected graph connected components problem. We could use recursion with a depth-first search. To avoid using the run-time stack, we could use a queue to perform a breadth-first search. Both of these run in $O(E + V)$ time. I would recommend in general to use the BFS.

4.1.2 Union-Find (Disjoint Set Union)

The union-find data structure is another way for us to solve the connected components problem. Union-find is unique from the other search techniques in that it can process input as it is presented, edge by edge. This also means it is possible to add more edges at the end, therefore changing the graph, while still running quickly. An algorithm that works like this is an *online algorithm*, while an algorithm that requires all the input data presented at the beginning is an *offline algorithm*.

A natural idea for solving the connected components problem is for each vertex to maintain a pointer to another vertex it's connected to, forming a *forest*, or collection of trees. To check whether two elements are in the same component, simply trace the tree up to the root by jumping up each pointer.

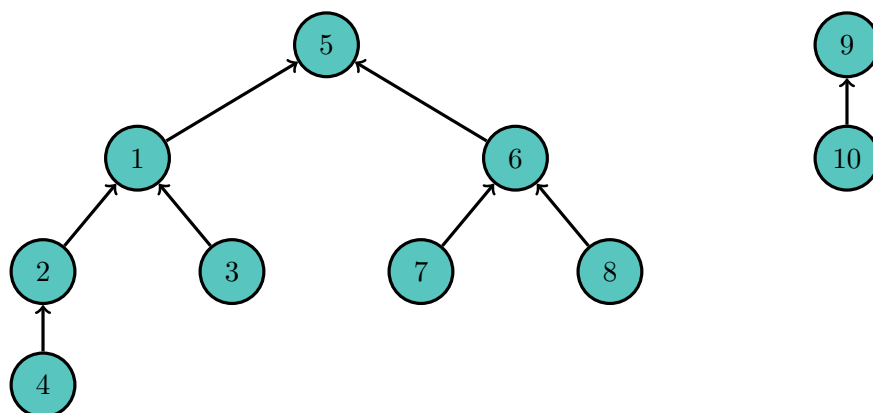


The idea of a pointer can easily be stored within an array.

-1	1	1	2	-1	5	6	6	-1	9
1	2	3	4	5	6	7	8	9	10

We want to support two operations: $find(v)$, which returns the root of the tree containing v , and $union(u, v)$, which merges the components containing u and v . This second operation is easy given the first; simply set the pointer of $find(u)$ to be $find(v)$.

$union(4, 6)$, unoptimized:

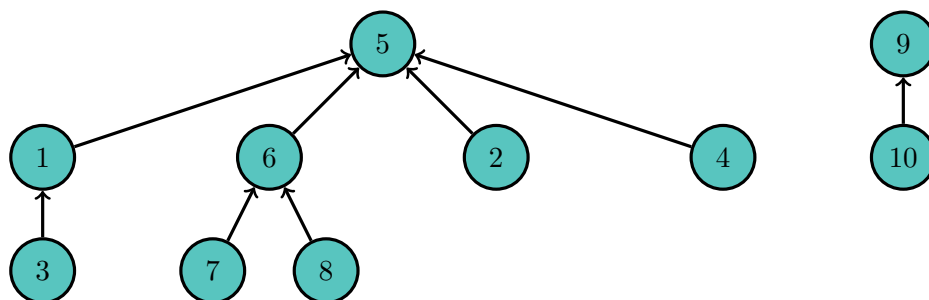


A problem quickly arises – the *find* operation threatens to become linear. There are two simple things we can do to optimize this.

The first is to always add the shorter tree to the taller tree, as we want to minimize the maximum height. An easy heuristic for the height of the tree is simply the number of elements in that tree. We can keep track of the size of the tree with a second array. This heuristic is obviously not perfect, as a larger tree can be shorter than a smaller tree, but it turns out with our second optimization that this problem doesn't matter.

The second fix is to simply assign the pointer associated with v to be $\text{find}(v)$ at the end of the *find* operation. We can design $\text{find}(v)$ to recursively call *find* on the pointer associated with v , so this fix sets pointers associated with nodes along the entire chain from v to $\text{find}(v)$ to be $\text{find}(v)$. These two optimizations combined make the *union* and *find* operations $O(\alpha(V))$, where $\alpha(n)$ is the inverse Ackermann function, and for all practical values of n , $\alpha(n) < 5$.

find(4), optimized:



Algorithm 1 Union-Find

```

function FIND( $v$ )
    if  $v$  is the root then
        return  $v$ 
     $parent(v) \leftarrow \text{FIND}(parent(v))$ 
    return  $parent(v)$ 

function UNION( $u, v$ )
     $uRoot \leftarrow \text{FIND}(u)$ 
     $vRoot \leftarrow \text{FIND}(v)$ 
    if  $uRoot = vRoot$  then
        return
    if  $size(uRoot) < size(vRoot)$  then
         $parent(uRoot) \leftarrow vRoot$ 
         $size(vRoot) \leftarrow size(uRoot) + size(vRoot)$ 
    else
         $parent(vRoot) \leftarrow uRoot$ 
         $size(uRoot) \leftarrow size(uRoot) + size(vRoot)$ 

```

4.1.3 Tarjan

Tarjan's algorithm for strongly connected components

4.2 Shortest Path

A classic. Assign nonnegative weights to each of the edges, where the weight of the edge (u, v) represents the distance from u to v . This graph can be either directed or undirected.

4.2.1 Dijkstra

Dijkstra's algorithm solves the single-source shortest path problem. From any vertex, we can compute the shortest path to each of the remaining vertices in the graph. The two formulations of Dijkstra's algorithm run in $O(V^2)$ or $O(E \log V)$ time, whichever one suits us better. Note that it is possible to do better than $O(E \log V)$ using a Fibonacci heap. The former works nicely on dense graphs, as $E \approx V^2$, while the latter works better on sparse graphs, as $E \approx V$.

For every vertex v in the graph, we keep track of the shortest known distance $dist(v)$ from the source to v , a boolean $visited(v)$ to keep track of which nodes we "visited," and a pointer to the previous node in the shortest known path $prev(v)$ so that we can trace the shortest path once the algorithm finishes.

Dijkstra iteratively "visits" the next nearest vertex, updating the distances to that vertex's neighbors if necessary. Therefore, at any step, we have the first however-many nearest vertices to the source, which we call "visited" and for which the shortest path is known. We also have the shortest path to all the remaining vertices that stays within the "visited" vertices besides for the very last edge, if such a path exists. We claim that the known distance to the closest vertex that

has not yet been visited is the shortest distance. We can then “visit” that vertex. It shouldn’t be hard to prove that this algorithm indeed calculates the shortest path.

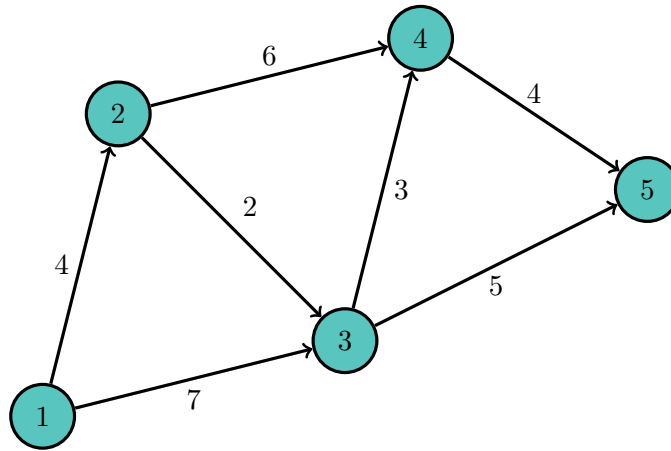
The $O(V^2)$ implementation immediately follows.

Algorithm 2 Dijkstra

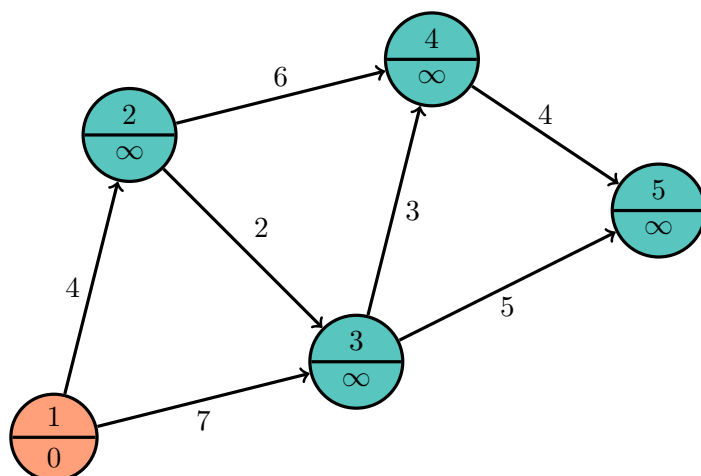
```

for all vertices  $v$  do
   $dist(v) \leftarrow \infty$ 
   $visited(v) \leftarrow 0$ 
   $prev(v) \leftarrow -1$ 
 $dist(src) \leftarrow 0$ 
while  $\exists v$  s.t.  $visited(v) = 0$  do
   $v \equiv v$  s.t.  $visited(v) = 0$  with min  $dist(v)$ 
   $visited(v) \leftarrow 1$ 
  for all neighbors  $u$  of  $v$  do
    if  $visited(u) = 0$  then
       $alt \leftarrow dist(v) + weight(v, u)$ 
      if  $alt < dist(u)$  then
         $dist(u) \leftarrow alt$ 
         $prev(u) \leftarrow v$ 

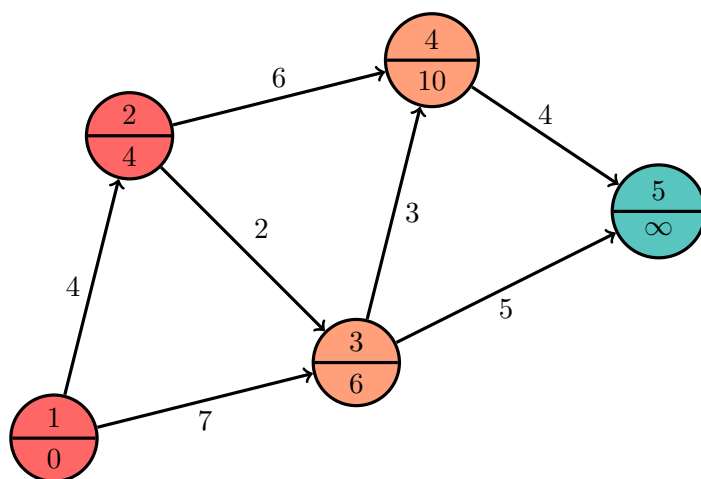
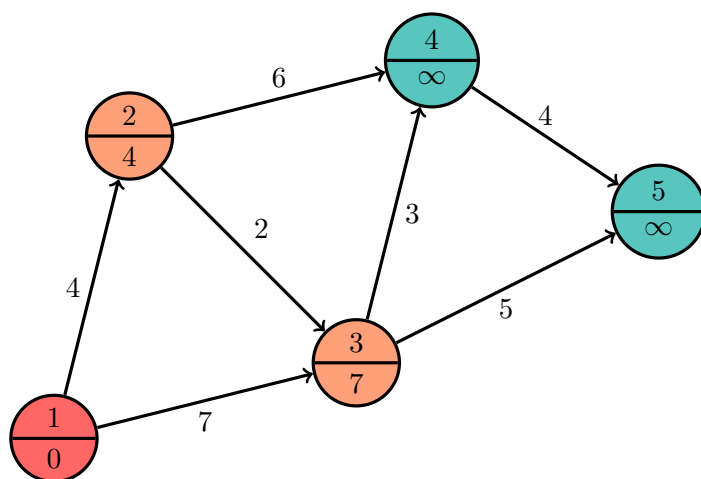
```

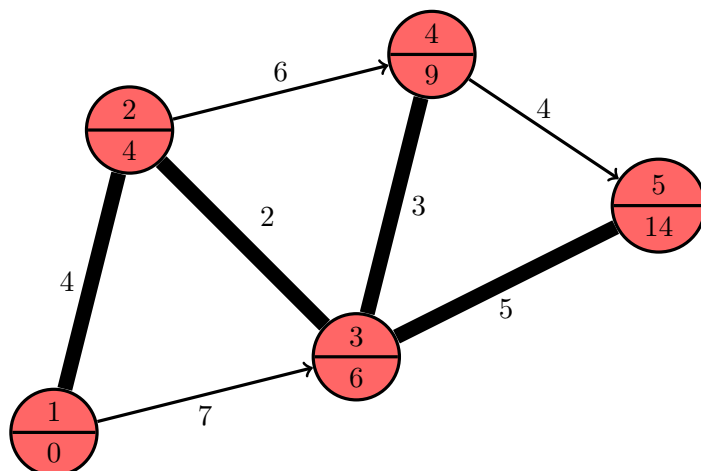
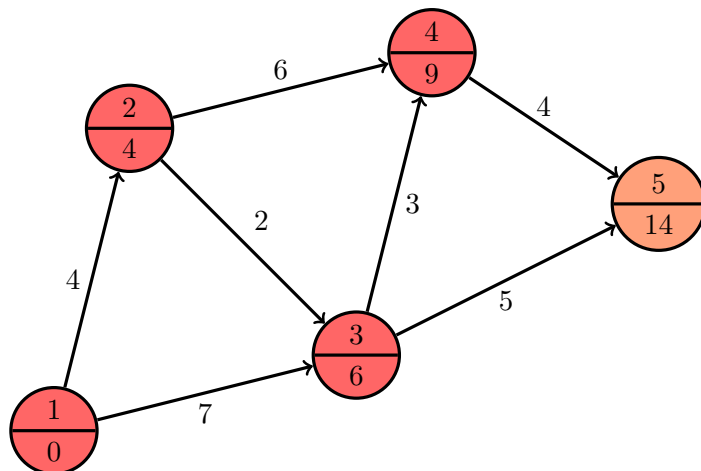
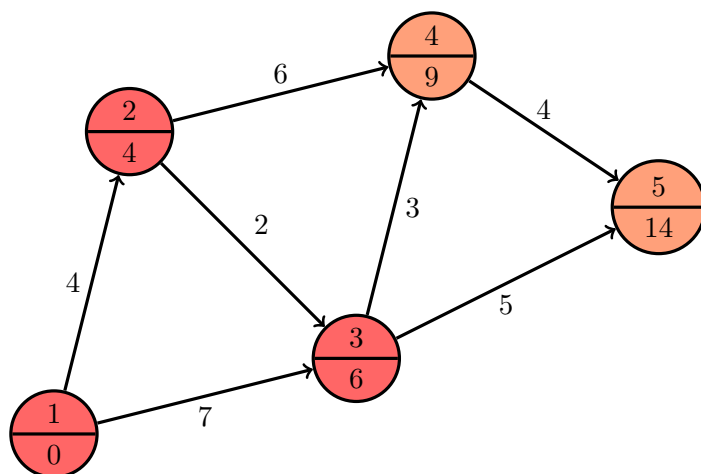


Let’s run Dijkstra’s algorithm on the above graph with vertex 1 as the source. We first set all the distances besides the source to be ∞ .



Now, we continue choosing the closest unvisited node, mark it as visited, and update its neighbors.





The slow part of the $O(V^2)$ formulation is the linear search for the vertex v with the minimum $dist(v)$. We happen to have a data structure that resolves this problem – a binary heap. The main problem with using the standard library heap is having repeated vertices in the heap. We could just ignore this problem and discard visited vertices as they come out of the heap. Alternatively, we

could choose never to have repeated vertices in the heap. To do this, we need to be able to change the value of the distances once they are already in the heap, or *decrease-key*. This is a pretty simple function to add, however, if you have a heap already coded. Either way, we achieve $O(E \log V)$, as we do $E + V$ updates to our heap, each costing $O(V)$.

4.2.2 Floyd-Warshall

Dijkstra is nice when we are dealing with edges with nonnegative weights and are looking for the distances from one vertex to all the others. Floyd-Warshall solves the shortest path problem for all pairs of vertices in $O(V^3)$ time, which is faster than V single-source Dijkstra runs on a dense graph. Floyd-Warshall works even if some edge weights are negative but not if the graph has a negative cycle.

Algorithm 3 Floyd-Warshall

```

for all vertices  $v$  do
     $dist(v, v) = 0$ 
for all edges  $(u, v)$  do
     $dist(u, v) = weight(u, v)$ 
for all vertices  $k$  do
    for all vertices  $i$  do
        for all vertices  $j$  do
            if  $dist(i, j) > dist(i, k) + dist(k, j)$  then
                 $dist(i, j) \leftarrow dist(i, k) + dist(k, j)$ 

```

4.2.3 Bellman-Ford

Bellman-Ford is a single-source $O(VE)$ shortest path algorithm that works when edge weights can be negative. It is preferable to Floyd-Warshall when the graph is sparse and we only need the answer for one source. Like Floyd-Warshall, the algorithm fails if the graph contains a negative cycle, but the algorithm is still useful for detecting negative cycles.

The idea here is the shortest path, assuming no negative cycles, has length at most $V - 1$.

Algorithm 4 Bellman-Ford

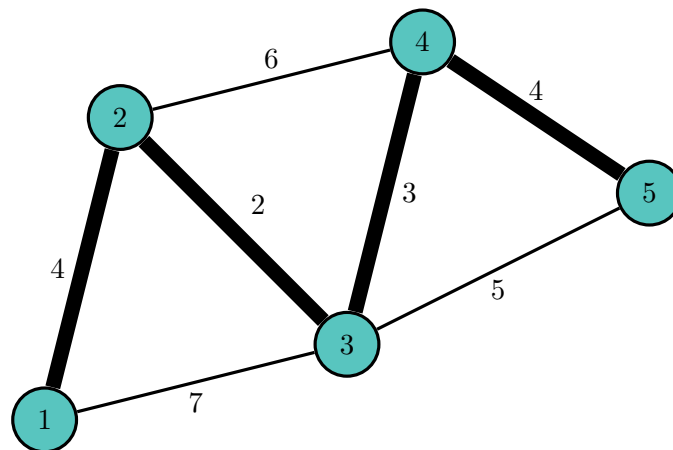
```

for all vertices  $v$  do
     $dist(v) \leftarrow \infty$ 
     $prev(v) \leftarrow -1$ 
 $dist(src) \leftarrow 0$ 
for  $i \equiv 1, V - 1$  do
    for all edges  $(u, v)$  do
        if  $dist(u) + weight(u, v) < dist(v)$  then
             $dist(v) \leftarrow dist(u) + weight(u, v)$ 
             $prev(v) \leftarrow u$ 
for all edges  $(u, v)$  do ▷ check for negative cycles
    if  $dist(u) + weight(u, v) < dist(v)$  then
        negative cycle detected

```

4.3 Minimum Spanning Tree

Consider a connected, undirected graph. A *spanning tree* is a subgraph that is a tree and contains every vertex in the original graph. A *minimum spanning tree* is a spanning tree such that the sum of the edge weights of the tree is minimized. Finding the minimum spanning tree uses many of the same ideas discussed earlier.



4.3.1 Prim

Prim's algorithm for finding the minimum spanning tree is very similar to Dijkstra's algorithm for finding the shortest path. Like Dijkstra, it iteratively adds a new vertex at a time to build a tree. The only difference is $dist(v)$ stores the shortest distance from *any* visited node instead of the source.

Algorithm 5 Prim

```

for all vertices  $v$  do
     $dist(v) \leftarrow \infty$ 
     $visited(v) \leftarrow 0$ 
     $prev(v) \leftarrow -1$ 
 $dist(src) \leftarrow 0$ 
while  $\exists v$  s.t.  $visited(v) = 0$  do
     $v \equiv v$  s.t.  $visited(v) = 0$  with  $\min dist(v)$ 
     $visited(v) \leftarrow 1$ 
    for all neighbors  $u$  of  $v$  do
        if  $visited(u) = 0$  then
            if  $weight(v, u) < dist(u)$  then
                 $dist(u) \leftarrow weight(v, u)$ 
                 $prev(u) \leftarrow v$ 

```

The proof of correctness is left as an exercise. The complexity of this algorithm depends on how the minimum unvisited vertex is calculated. Using the same approaches as Dijkstra, we can achieve $O(V^2)$ or $O(E \log V)$.

4.3.2 Kruskal

While Prim greedily adds vertices to the tree, Kruskal's algorithm greedily adds edges. It iterates over all the edges, sorted by weight. We need to watch out for adding a cycle, breaking the tree structure, which means we need to keep track of each vertex's connected component. If an edge connects two vertices from the same connected component, we don't want to add it to our tree. However, we have a union-find algorithm that works perfectly for this.

Algorithm 6 Kruskal

```

for all edges  $(u, v)$  in sorted order do
    if  $FIND(u) \neq FIND(v)$  then
        add  $(u, v)$  to spanning tree
         $UNION(u, v)$ 

```

This algorithm requires a sort of the edges and thus has complexity $O(E \log E) = O(E \log V)$.

4.4 Eulerian Tour

An *Eulerian tour* of a graph is a path that traverses every edge exactly once. If the tour ends exactly where it started, it is called an *Eulerian circuit*. A graph has an Eulerian circuit if it is connected and every vertex has even degree. A graph has an Eulerian path if it is connected and all vertices but exactly two have even degrees. The mathematical proofs for these graph properties hinge on the idea that removing a cycle from the graph maintains the Eulerian property. We construct an Eulerian tour by appealing to this idea.

Algorithm 7 Eulerian Tour

```

function FINDTOUR( $v$ )
  while  $v$  has a neighbor  $u$  do
    delete edge  $(v, u)$ 
    FINDTOUR( $u$ ) ▷ FINDTOUR( $u$ ) must trace a circuit back to  $v$ 
  add  $v$  to tour

```

It is not preferable to use the run-time stack; we can use our own stack if necessary.

If the graph contains an Eulerian circuit, we call this function on any vertex we like. If it contains an Eulerian path, we call this function on one of the vertices with odd degree.

4.5 Maximum Flow

We are given a directed graph with weighted edges, a source node, and a sink node. A *flow* is sent from the source to the sink. Each edge weight represents the maximum capacity of that edge. For every node besides the source and the sink node, total flow in is equal to total flow out. We can think of a flow network as a series of pipes through which water travels from an entrance to an exit in the network. The edge capacities represent pipe thickness. At any node, the total rate at which water enters the node must equal the total rate at which it exits the node, and along any path, the rate at which water flows is bottlenecked by the thinnest pipe.

More formally, for a graph $G(V, E)$, where $c(u, v)$ represents the capacity of the edge from u to v , the flow $f(u, v)$ from a source s to a sink t satisfies

$$\begin{aligned}
 f(u, v) &\leq c(u, v) \forall (u, v) \in E \\
 f(u, v) &= -f(v, u) \forall (u, v) \in E \\
 \sum_{v \in V} f(u, v) &= 0 \forall u \in V \setminus \{s, t\} \\
 \sum_{v \in V} f(s, v) &= \sum_{v \in V} f(v, t) = |f|,
 \end{aligned}$$

where $|f|$ represents the total flow from the source to the sink.

A *maximum flow* is one that maximizes the total flow $|f|$ from s to t , or in our example, maximizes the rate at which water can flow through our network.

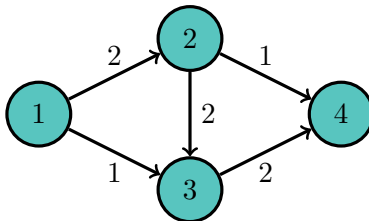
We'll also define the residual capacity $c_f(u, v) = c(u, v) - f(u, v)$. Note that $c_f(u, v) \geq 0$ by the conditions imposed on f . The residual capacity of an edge represents how much capacity is left after a certain amount of flow has already been sent. We therefore have the residual graph $G_f(V, E_f)$, where E_f is the graph of residual edges, or all edges $(u, v) \in V^2$ satisfying $c_f(u, v) > 0$.

A natural approach to “solving” this problem would be to simply greedily add flow.

Find a path from the source to the sink in which all the edges have positive weight in the residual graph. Send flow along this path; that is, find the max flow across this path, which is the minimum weight of any edge on this particular path. Call this value *cap*. Then subtract *cap* from the residual

capacity of every edge in the path. We repeat, and this is guaranteed to terminate since on any given move, we remove an edge from our residual graph.

What is wrong with our greedy approach? Consider the following graph:



The max flow from vertex 1 to vertex 4 is 3, but greedy gives only 2. This is because the best possible single path from the source to the sink may not be included the best possible overall flow.

4.5.1 Ford-Fulkerson

We somehow need a way to fix the inclusion of any suboptimal paths in our greedy approach, or to “send flow back” in case we sent it through a suboptimal path. We do this by introducing the *reverse edge* to our residual graph.

Find a path from the source to the sink in which all the edges have positive weight in the residual graph. Find the max flow across this path, which is the minimum weight of any edge on this particular path. Call this value *cap*. Then subtract *cap* from the residual capacity of every edge along the path and *increment the residual capacity of the reverse edge* (the edge connecting the same two vertices but running in the opposite direction) by *cap*. We call this operation on the path *augmenting the path*. We simply choose an augmenting path until no such paths exist.

Algorithm 8 Ford-Fulkerson

```

function AUGMENTPATH(path  $p = \{v_i\}_{i=1}^m$ , where  $(v_i, v_{i+1}) \in E_f$ ,  $v_1 = s$ ,  $v_m = t$ )
   $cap \leftarrow \min_{i=1}^{m-1} (c_f(v_i, v_{i+1}))$ 
  for  $i \equiv 1, m-1$  do
     $f(v_i, v_{i+1}) \leftarrow f(v_i, v_{i+1}) + cap$ 
     $c_f(v_i, v_{i+1}) \leftarrow c_f(v_i, v_{i+1}) - cap$ 
     $f(v_{i+1}, v_i) \leftarrow f(v_{i+1}, v_i) - cap$ 
     $c_f(v_{i+1}, v_i) \leftarrow c_f(v_{i+1}, v_i) + cap$ 
    ▷ incrementing reverse edge
function MAXFLOW( $G(V, E)$ ,  $s, t \in V$ )
  for all  $(u, v) \in V^2$  do
     $f(u, v) \leftarrow 0$ 
     $c_f(u, v) \leftarrow c(u, v)$ 
   $|f| \leftarrow 0$ 
  while  $\exists p = \{v_i\}_{i=1}^m$ , where  $(v_i, v_{i+1}) \in E_f$ ,  $v_1 = s$ ,  $v_m = t$  do
     $cap \leftarrow \min_{i=1}^{m-1} (c_f(v_i, v_{i+1}))$ 
     $|f| \leftarrow |f| + cap$ 
    AUGMENTPATH( $p$ )
  return  $|f|$ 

```

The difference between this algorithm and the greedy approach from earlier is that the paths we now allow may run along a reverse path, essentially undoing any suboptimal flow from earlier. These more general paths in our residual graph are called *augmenting paths*.

This algorithm is guaranteed to terminate for graphs with integral weights. Its performance is bounded by $O(Ef)$, where f is the maximum flow and E is the number of edges, as finding a path from s to t takes $O(E)$ and increments the total flow f by at least 1. The concept of removing edges can't be used to produce a stricter bound because while an edge in one direction may be removed from the residual graph, doing so creates an edge in the other direction.

In its crudest form, Ford-Fulkerson does not specify on which path to push flow if multiple paths exist. It simply states that as long as such a path exists, push flow onto it. In addition to being slow, Ford-Fulkerson, as it is stated, is not guaranteed to terminate for graphs with non-integral capacities. In fact, it might not even converge to the maximum flow for irrational capacities. However, these problems can be fixed by simply specifying how the algorithm chooses the next path on which to push flow. Nonetheless, the Ford-Fulkerson algorithm is formulated beautifully mathematically and such is useful from a math perspective, as we will see with the Max-Flow Min-Cut Theorem.

4.5.2 Max-Flow Min-Cut Theorem

On a graph $G(V, E)$, an s - t cut $C = (S, T)$ splits the partitions V into S and T satisfying $s \in S$ and $t \in T$. The *cut-set* of C is the set of edges $\{(u, v) \in E : u \in S, v \in T\}$. The *capacity* of an s - t cut is given by

$$c(S, T) = \sum_{(u,v) \in S \times T} c(u, v).$$

A *minimum cut* is an s - t cut that minimizes $c(S, T)$.

A cut represents a set of edges that, once removed, separates s from t . A minimum cut is therefore a set of edges that does this but minimizes the total capacity of the edges necessary to disconnect s from t .

The max-flow min-cut theorem states that the maximum value of any s - t flow is equal to the minimum capacity of any s - t cut.

First, the capacity of any cut must be at least the total flow. This is true by contradiction. Any path from s to t has an edge in the cut-set, so therefore any flow from s to t is upper bounded by the capacity of the cut. Therefore, we need only construct one flow and one cut such that the capacity of the cut is equal to the flow.

We consider the residual graph G_f produced at the completion of the Ford-Fulkerson augmenting path algorithm.¹ Let the set S be all nodes reachable from s and T be $V \setminus S$. We wish to show that $C = (S, T)$ is a cut satisfying $|f| = c(S, T) = \sum_{(u,v) \in S \times T} c(u, v)$. This is true when the following is satisfied:

1. All edges $(u, v) \in S \times T$ are fully saturated by the flow. That is, $c_f(u, v) = 0$.

¹If you're concerned that the Ford-Fulkerson algorithm will never terminate, there always exists a sequence of paths chosen such that it will. Edmonds-Karp is one example that always terminates.

2. All reverse edges $(v, u) \in T \times S$ have zero flow. That is, $f(v, u) = 0$, or $c_f(v, u) = c(v, u)$.

The first condition is true by the way we constructed S and T , as if there existed a (u, v) where $c_f(u, v) > 0$, then v is accessible to s and ought to have been in S .

The second condition is true by the way the Ford-Fulkerson algorithm constructed reverse edges. If net flow was sent from v to u , then a reverse edge was constructed from u to v , so again, v is accessible to s , which is a contradiction.

Therefore, we have the flow

$$|f| = \sum_{(u,v) \in S \times T} c(u, v) - \sum_{(v,u) \in T \times S} 0 = c(S, T),$$

so we constructed a flow and a cut such that the flow $|f|$ is equal to the cut capacity $c(S, T)$, and we are done.

4.5.3 Edmonds-Karp

Edmonds-Karp is a refinement of Ford-Fulkerson. As stated earlier, Ford-Fulkerson is limited by not specifying on which path to push flow. There are many algorithms that resolve this issue in different ways; Edmonds-Karp fixes this by simply choosing the augmenting path of shortest unweighted length. This can be done easily using a BFS.

Algorithm 9 Edmonds-Karp

function CHOOSEPATH($G_f(V, E_f), s, t \in V$) ▷ BFS
 $visited(v)$ denotes v has been added to queue
 $prev(v)$ denotes vertex preceding v in BFS
 push s on queue q
 $visited(s) \leftarrow 1$
 while q is not empty **do**
 $u \leftarrow$ top of q
 for all neighbors v of u in G_f where $visited(v) = 0$ **do**
 push v on q
 $visited(v) \leftarrow 1$
 $prev(v) \leftarrow u$
 pointer $curr \leftarrow t$
 while $curr \neq s$ **do**
 add $curr$ to beginning of path p
 $curr \leftarrow prev(curr)$
 add s to beginning of p
 return p

function MAXFLOW($G(V, E), s, t \in V$)
 for all $(u, v) \in V^2$ **do**
 $f(u, v) \leftarrow 0$
 $c_f(u, v) \leftarrow c(u, v)$
 $|f| \leftarrow 0$
 while t can be reached from s **do**
 $p \leftarrow$ CHOOSEPATH(G_f, s, t)
 $cap \leftarrow \min_{i=1}^{m-1} (c_f(v_i, v_{i+1}))$
 $|f| \leftarrow |f| + cap$
 AUGMENTPATH(p)
 return $|f|$

The BFS is clearly $O(E)$. To complete our analysis, we must somehow bound the number of times we need to perform the BFS. To do this, we'll look at what pushing flow on a path does to our residual graph; in particular, how it affects our BFS traversal tree. Note that each vertex is on some level i in the BFS tree, characterized by the distance from the source s . For example, $L_0 = \{s\}$, L_1 contains all the neighbors of s , L_2 contains neighbors of neighbors not in L_0 or L_1 , and so on.

We first claim that the level of any vertex in the graph is nondecreasing following an augment on a path p . If the augment saturates an edge, it may remove it from G_f , which cannot decrease the distance of any vertex from s . If the augment creates an edge $e = (u, v)$, that means we sent flow from v to u on the path p . Therefore, if v was originally level i , u must have been level $i + 1$. The level of u does not change by adding (u, v) , and the level of v can either be i or $i + 2$, depending on whether edge (v, u) was deleted in the process. Either way, the level of all vertices is nondecreasing.

Now consider the bottleneck edge $e = (u, v)$ of an augmenting path p , where the level of u is i and the level of v is $i + 1$. The push operation deletes the edge e , but the level of v must stay at least $i + 1$. Now for the edge e to reappear in the graph G_f , flow must have been sent on the reverse

edge $e' = (v, u)$ on some augmenting path p' . But on path p' , u comes after v , which must be at least level $i + 1$. Therefore, u must be at least level i . But since the maximum level of a node that is connected to s is $V - 1$, an edge e can only be chosen as the bottleneck edge $\frac{V}{2}$ times, or $O(V)$.

There are E edges, each of which can be the bottleneck edge for $O(V)$ different augmenting paths, each of which takes $O(E)$ to process. Therefore, the Edmonds-Karp algorithm runs in $O(VE^2)$.

4.5.4 Dinic

Dinic's algorithm is another refinement of Ford-Fulkerson.

4.5.5 Push-Relabel

Unfortunately, even the much-improved bounds of the $O(VE^2)$ Edmonds-Karp are admittedly bad. While the push-relabel method for solving the max flow problem does not have the fastest theoretical bounds, two of its implementations have complexities $O(V^3)$ and $O(V^2\sqrt{E})$ and are among the fastest in practice.

Generic Push-Relabel

Ford-Fulkerson and its variants all deal with global augmentations of paths from s to t . Push-relabel takes a different perspective, introducing the concept of a *preflow* and a *height* to make local optimizations that ultimately result in the maximum flow.

A preflow maintains the same properties as a flow but modifies the conservation of flow condition. Instead of total flow in equalling total flow out, flow in must be at least, and therefore can exceed, flow out. We denote the difference between flow in and flow out as the *excess* $e(v)$.

$$\begin{aligned} f(u, v) &\leq c(u, v) \forall (u, v) \in E \\ f(u, v) &= -f(v, u) \forall (u, v) \in E \\ e(v) &= \sum_{u \in V} f(u, v) \geq 0 \forall v \in V \setminus \{s\} \\ e(s) &= \infty. \end{aligned}$$

The definitions of the residual capacity $c_f(u, v)$, edge set E_f , and graph G_f are the same as they were defined before, except with a preflow f instead of a normal flow.

We call a vertex $v \in V \setminus \{s, t\}$ *active* if $e(v) > 0$. Therefore, a vertex besides the source or sink is active if more flows into the vertex than flows out. s and t are *never* active. A preflow with no active vertices is simply a flow, at which point the excess of the sink $e(t)$ represents the value $|f|$ of the flow.

We can *push* flow from a node u to a node v by moving as much of the excess $e(u)$ to v as the capacity of the edge $c_f(u, v)$ will allow.

```

function PUSH(edge  $(u, v)$ )
     $\delta \leftarrow \min(e(u), c_f(u, v))$ 
     $f(u, v) \leftarrow f(u, v) + \delta$ 
     $f(v, u) \leftarrow f(v, u) - \delta$ 
     $e(u) \leftarrow e(u) - \delta$ 
     $e(v) \leftarrow e(v) + \delta$ 

```

$\triangleright c_f(u, v) = c(u, v) - f(u, v)$

The idea of the push-relabel algorithm is to first push as much preflow as possible through local optimizations in the direction the sink. When a node can no longer push flow to the sink, it pushes the excess back towards the source to turn the preflow into a flow.

However, the difficulty here lies in establishing this sense of “direction” from the source to the sink. Remember that we simply push preflow along a single edge in the graph at a time, not along a whole path. Moving flow from the source to the sink along a path that goes from the source to the sink is easy; moving flow from the source to the sink through local pushes without the knowledge of the graph structure as a whole is indeed a much harder problem.

To resolve this issue, we introduce a *label* to each of the nodes. The label $h(u)$ represents the “height” of u . In real life, water flows from higher to lower ground. We want s to represent that high ground and t to represent the low ground. As we push preflow from s to t , vertices along the way represent height values between those of s and t . However, eventually we have to push preflow back to handle both excesses in flow and suboptimal previous pushes, à la Ford-Fulkerson, but this contradicts the concept of height as we can’t flow both downhill and uphill. Therefore, we’ll need to be able to *relabel* a node, changing the height $h(u)$ to something that allows preflow to flow back towards s . We will relabel h in a systematic way that allows us to direct the preflow through the graph.

For this labeling to be useful for us, we’ll need to impose some more constraints that must be satisfied no matter how we change the graph G_f or the height function h .

$$\begin{aligned}
 h(u) &\leq h(v) + 1 \forall (u, v) \in E_f \\
 h(s) &= |V| \\
 h(t) &= 0.
 \end{aligned}$$

What does this mean? For our algorithm, we can push preflow along the edge from u to v only if $c_f(u, v) > 0$ and $h(u) > h(v)$, so $h(u) = h(v) + 1$. We call such an edge $(u, v) \in E_f$ *admissible*. Furthermore, for all vertices v that can reach t in E_f , $h(v)$ represents the lower bound for the length of any unweighted path from v to t in G_f , and for all vertices that cannot reach t , then $h(v) - |V|$ is a lower bound for the unweighted distance from s to v .

t will always represent the lowest node, so $h(t) = 0$ is a natural constraint. We’ll first set the preflow values of all vertices v that can be immediately reached from s to $c(s, v)$, saturating all the out-edges of s . For any vertex v from which t can be reached, $h(v)$ represents the lower bound of the unweighted distance to t from v in the residual graph.

We want $h(s)$ to be a number large enough that will indicate that s has been disconnected to t , as we have already sent as much preflow possible from s in the direction of t by saturating all outgoing edges. Therefore, setting $h(s) = |V|$ is also natural. Since $h(s)$ represents the lower bound of the distance from s to t in G_f , and there are no paths from s to t in the residual graph, $|V|$ is a natural choice, since the longest possible path is $|V| - 1$.

Furthermore, we don't want any preflow sent back to s from a vertex v unless it is impossible to send any more preflow from v to t . If preflow is pushed from v to s , then $h(v) = |V| + 1$. If there existed a path v to t such that every edge is admissible, the path must have $|V| + 2$ vertices. This is true because for any two consecutive vertices v_i, v_{i+1} in the path, $h(v_i) = h(v_{i+1}) + 1$, but no path can have $|V| + 2$ distinct vertices.

This leads to the fact that the only nodes that can possibly continue to contribute to the final flow are active vertices v for which $h(v) < |V|$. A node with height at least $|V|$ does not have a valid path to t , and a node that is not active doesn't have any excess flow to push.

Now that I've explained the general idea behind the labeling constraints, it's time to actually describe what our relabeling process is. At first, the labels of all vertices besides the source start at 0. We only relabel a node v if it is active (therefore, it has excess flow it needs to push; $e(u) > 0$) but has no admissible out-edges in G_f (so it has no adjacent vertex on which it can push that excess flow). If a node has no admissible out-edges in G_f , every neighbor of u has a height label at least equal to $h(u)$. When we relabel a node, we always then *increase* the value of $h(u)$ to the least value where it can push flow onto another node.

function RELABEL(vertex u)

$$h(u) \leftarrow \min_{v|(u,v) \in E_f} (h(v)) + 1 \quad \triangleright (u, v) \in E_f \iff c_f(u, v) = c(u, v) - f(u, v) > 0$$

Since we take the minimum height of all neighbors in the graph, we first try adjusting the height of u so that we can push flow from u to its neighbors that can possibly still reach t ; that is, neighbors v satisfying $h(v) < |V|$. Once we try all these neighbors, we then increase the height of u to begin to push flow back towards s . We can always find such an edge, as any preflow pushed onto u must have also incremented the reverse edge from u back towards s .

Note that neither pushing on a admissible edge nor relabeling a vertex with no admissible out-edges changes the fact that h remains a valid labeling function.

The generic push-relabel algorithm simply pushes and relabels vertices until there are no active vertices and the preflow becomes a flow. This algorithm works because throughout the process, h remained a valid height function, and at the end, the preflow was converted into a flow. Since $h(s) = |V|$ and $h(t) = 0$, there is no augmenting path from s to t , so our flow is maximal.

Algorithm 10 Push-Relabel (Generic)

```

function PUSH(edge  $(u, v)$ )
  if  $e(u) > 0$  and  $h(u) = h(v) + 1$  then                                 $\triangleright$  push condition
     $\delta \leftarrow \min(e(u), c_f(u, v))$                                  $\triangleright c_f(u, v) = c(u, v) - f(u, v)$ 
     $f(u, v) \leftarrow f(u, v) + \delta$ 
     $f(v, u) \leftarrow f(v, u) - \delta$ 
     $e(u) \leftarrow e(u) - \delta$ 
     $e(v) \leftarrow e(v) + \delta$ 

function RELABEL(vertex  $u$ )
  if  $u \neq s, t$  and  $e(u) > 0$  and  $h(u) \leq h(v) \forall v | (u, v) \in E_f$  then     $\triangleright$  relabel condition
     $h(u) \leftarrow \min_{v | (u, v) \in E_f} (h(v)) + 1$      $\triangleright (u, v) \in E_f \iff c_f(u, v) = c(u, v) - f(u, v) > 0$ 

function MAXFLOW( $G(V, E)$ ,  $s, t \in V$ )
  for all  $v \in V \setminus \{s\}$  do                                 $\triangleright$  initialize excess
     $e(v) \leftarrow 0$ 
   $e(s) \leftarrow \infty$ 
  for all  $(u, v) \in V^2$  do                                 $\triangleright$  initialize preflow
     $f(u, v) \leftarrow 0$ 
  for all neighbors  $v \neq s$  of  $s$  do                                 $\triangleright$  saturate edges to all neighbors of  $s$ 
     $f(s, v) \leftarrow c(s, v)$ 
     $f(v, s) \leftarrow -c(s, v)$ 
     $e(v) \leftarrow c(s, v)$ 
  for all  $v \in V \setminus \{s\}$  do                                 $\triangleright$  preflow now has valid height function; initialize height
     $h(v) \leftarrow 0$ 
   $h(s) \leftarrow |V|$ 
  while we can still PUSH or RELABEL do
    PUSH or RELABEL
  return  $e(t)$                                  $\triangleright e(t) = |f|$ 

```

We can argue that this algorithm runs in $O(V^2E)$, which is already an improvement from Edmonds-Karp. However, just as Ford-Fulkerson could be sped up by specifying which augmenting paths to choose, we can do the same with the push-relabel algorithm, speeding it up by specifying a systematic method to choose an edge to push or a vertex to relabel.

Discharge

We first describe an auxiliary operation. For each vertex u , we'll need a way to visit in- and out-neighbors of u in a static cyclic order. This is easy with just a pointer; for vertex u , we'll call that pointer $curr(u)$. When the pointer passes through every element in the list of neighbors, we'll just reset it back to the first element.

```

function DISCHARGE(vertex  $u$ )
  while  $e(u) > 0$  do                                ▷ perform an operation as long as  $u$  is active
    if  $curr(u)$  is at end of list of neighbors then
      RELABEL( $u$ )
      reset  $curr(u)$ 
    else
      if  $(u, curr(u))$  is an admissible edge then
        PUSH( $(u, curr(u))$ )
      else
        move  $curr(u)$  to next neighbor of  $u$ 

```

FIFO Selection

FIFO selection simply maintains a list of active vertices in a FIFO queue. We pop off the first vertex in the queue and discharge it, adding any newly-activated vertices to the end of the queue. This runs in $O(V^3)$.

Highest Label Selection

Highest label selection discharges the active vertex with the greatest height. This runs in $O(V^2\sqrt{E})$.

Improvements with Heuristics

Heuristics are meant to help relabel vertices in a smarter way. Bad relabelings are the slowest part of the algorithm, and improving the process can speed up max flow.

The *gap heuristic* takes advantage of “gaps” in the height function. Since a path of admissible edges consists of vertices whose heights decrease by exactly 1, the presence of a gap in height precludes the possibility of such a path. If there exists a value h' such that no vertex v exists such that $h(v) = h'$, then for every vertex v satisfying $h' < h(v) < |V|$, v has been disconnected from t , so we can immediately relabel $h(v) = |V| + 1$.

The *global relabeling heuristic* performs a backwards BFS from t every now and then to compute the heights of the vertices in the graph exactly.

Some dude on Codeforces² didn’t have much luck improving performance with the global relabeling heuristic. I’d suggest sticking to the gap heuristic only.

4.5.6 Other Problems with Flow Solutions

Bipartite Matching

Edge-Disjoint Paths

²<http://codeforces.com/blog/entry/14378>

Chapter 5

Computational Geometry

For actual geometry problems, and not graph theory problems hiding in the plane.

I'm too lazy to actually write this section right now so here are some useful links from the USACO Training Pages.

<https://www.dropbox.com/s/nqzk63bjby1iaq9/Computational%20Geometry.pdf?dl=0>

<https://www.dropbox.com/s/ykf65dk6sefb6zk/2-D%20Convex%20Hull.pdf?dl=0>

These essentially cover anything I would want to say in this chapter anyway, so I'll likely fill this chapter out last.

5.1 Basic Tools

Cross Product

Dot Product

\tan^{-1} , atan2

5.2 Formulas

5.2.1 Area

5.2.2 Distance

5.2.3 Configuration

5.2.4 Intersection

5.3 Convex Hull

Chapter 6

Strings

6.1 String Hashing

6.2 Rabin-Karp

6.3 Knuth-Morris-Pratt

https://activities.tjhsst.edu/sct/lectures/1415/stringmatching_10_3_14.pdf

6.4 Trie

6.5 Suffix Data Structures

6.5.1 Suffix Tree

6.5.2 Suffix Array

6.6 Aho-Corasick

Chapter 7

More Stuff

Here are some more tools to stuff in the toolbox. This chapter introduces more complex ideas that are useful for solving USACO Gold problems and beyond.

7.1 \sqrt{n} Bucketing

\sqrt{n} bucketing is a relatively straightforward idea – given n elements $\{a_i\}_{i=1}^n$ in a sequence, we group them into \sqrt{n} equal-sized buckets. The motivation for arranging elements like this is to support an operation called a *range query*.

Let's take a concrete example. Suppose we want to support two operations:

- *update*(i, x) – increment the value of a_i by x
- *query*(i, j) – return $\sum_{k=i}^j a_k$.

Suppose we simply stored the sequence in an array. *update* then becomes an $O(1)$ operation, but *query* is $O(n)$.

2	4	7	-5	3	6	-3	1	-2	-4	-6	2	8	6	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Another natural approach would be to store in a separate array the sum of the first i terms in the sequence for every index i , or store the *prefix sums*.

0	2	6	13	8	11	17	14	15	13	9	3	5	13	19	19	12
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Now *query* becomes an $O(1)$ operation, as we can simply subtract two elements in the array to answer a query. Unfortunately, *update* becomes $O(n)$, as changing the value of an element in the beginning of the sequence forces us to change almost all the values in the prefix sum array.

We can still use this idea, though... what we are looking for is some way to group values into sums such that we only need to change a small number of the sums to *update* and only require a small number of them to *query*.

This leads us directly to a \sqrt{n} bucketing solution. Let's group the 16 elements into 4 groups.

2	4	7	-5	3	6	-3	1	-2	-4	-6	2	8	6	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
8				7				-10				7			
[1, 4]				[5, 8]				[9, 12]				[13, 16]			

We'll keep track of the total sum of each group. Now, if we want to update a value, we need to change only two values – the value of that element in the original array and the total sum of the bucket it is in. When we query a range, we'll take advantage of the sum of the bucket when we can. Highlighted are the numbers we'll need for *query*(7, 15).

2	4	7	-5	3	6	-3	1	-2	-4	-6	2	8	6	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
8				7				-10				7			
[1, 4]				[5, 8]				[9, 12]				[13, 16]			

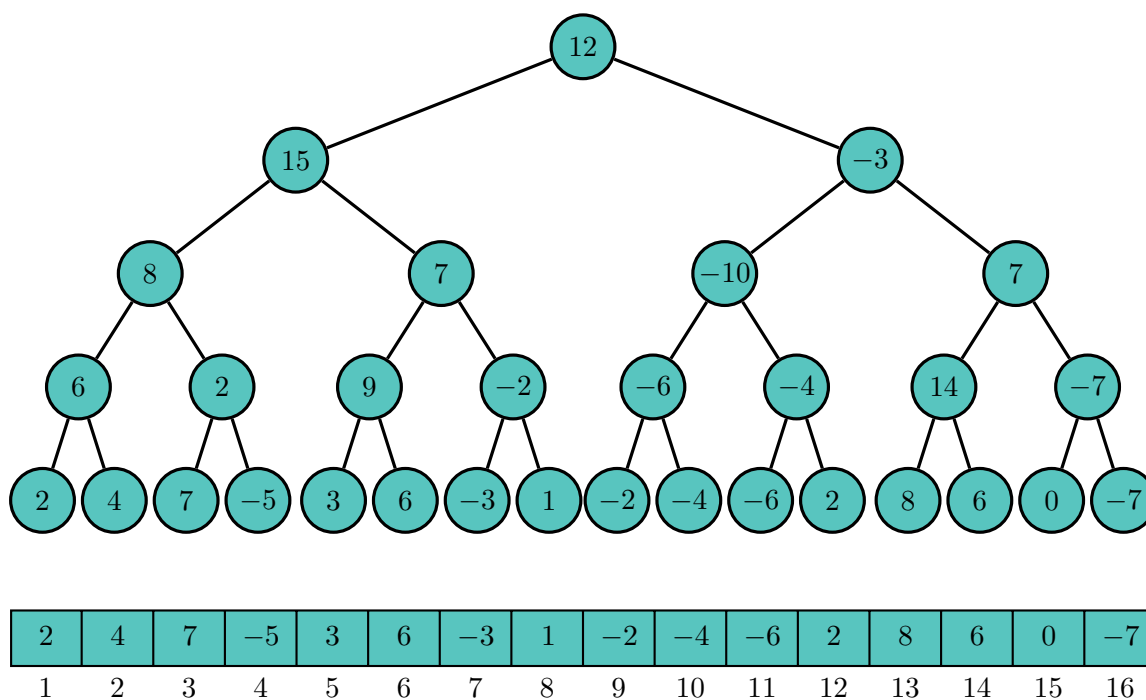
Querying requires access to at most \sqrt{n} bucket sums and $2(\sqrt{n} - 1)$ individual values. Therefore we have $O(\sqrt{n})$ query and $O(1)$ update. We are able to improve $O(\sqrt{n})$ update to $O(1)$ because of nice properties of the $+$ operator. This is not always the case for range queries: suppose, for instance, we needed to find the minimum element on a range.

It is often the case that $O(\sqrt{n})$ bounds can be improved to $O(\log n)$ using more complex data structures like segment trees and more complex ideas like 2^n jump pointers, both of which are covered in this chapter. These are, however, more complicated to implement and as such are often comparable in runtime in the contest environment. Steven Hao is notorious for using crude \sqrt{n} bucketing algorithms to solve problems that should have required tighter algorithm complexities. \sqrt{n} bucketing is a crude yet powerful idea; always keep it in the back of your mind.

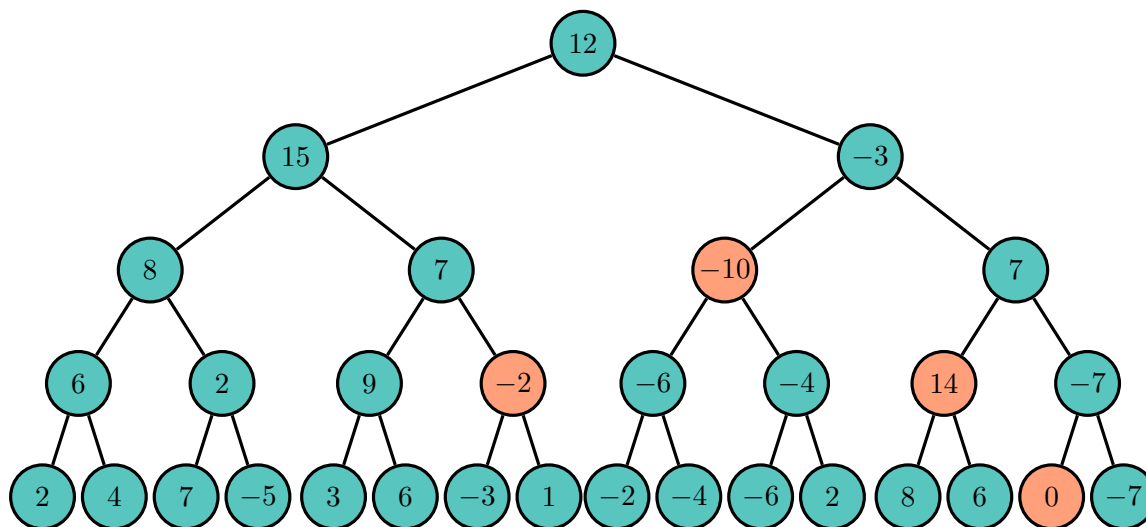
7.2 Segment Tree

It turns out for our specific sum problem that we can do as good as $O(\log n)$ with a *segment tree*, or *range tree*, or *augmented static BBST*. The essential idea is still the same – we want to group elements in some way that allows us to update and query efficiently.

As the name “tree” suggests, we draw inspiration from a binary structure. Let's build a tree on top of the array, where each node keeps track of the sum of the numbers associated with its children.

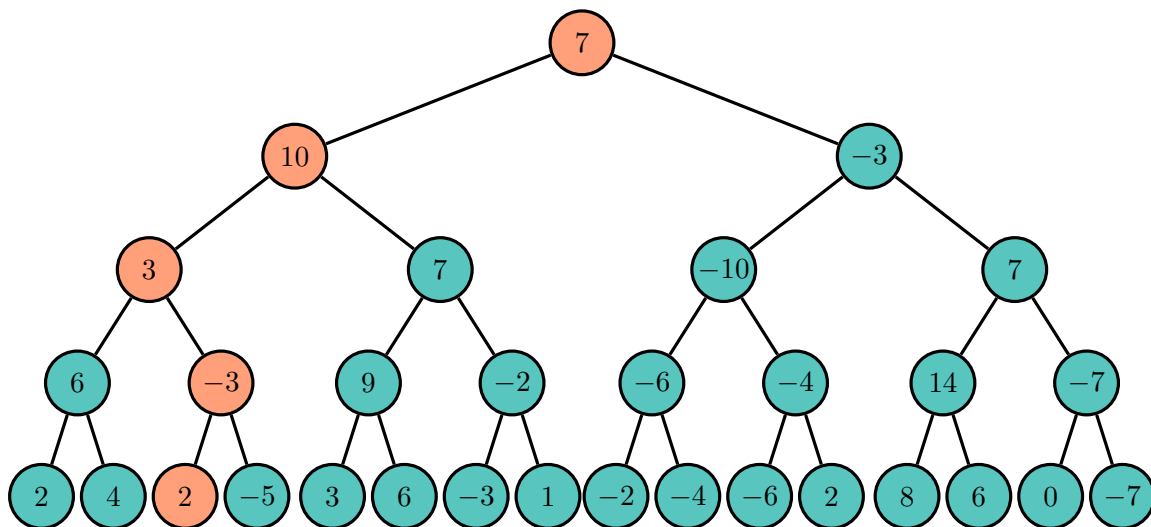


Highlighted are the nodes we'll need to access for $query(7, 15)$. Notice how the subtrees associated with each of these nodes neatly covers the entire range $[7, 15]$.



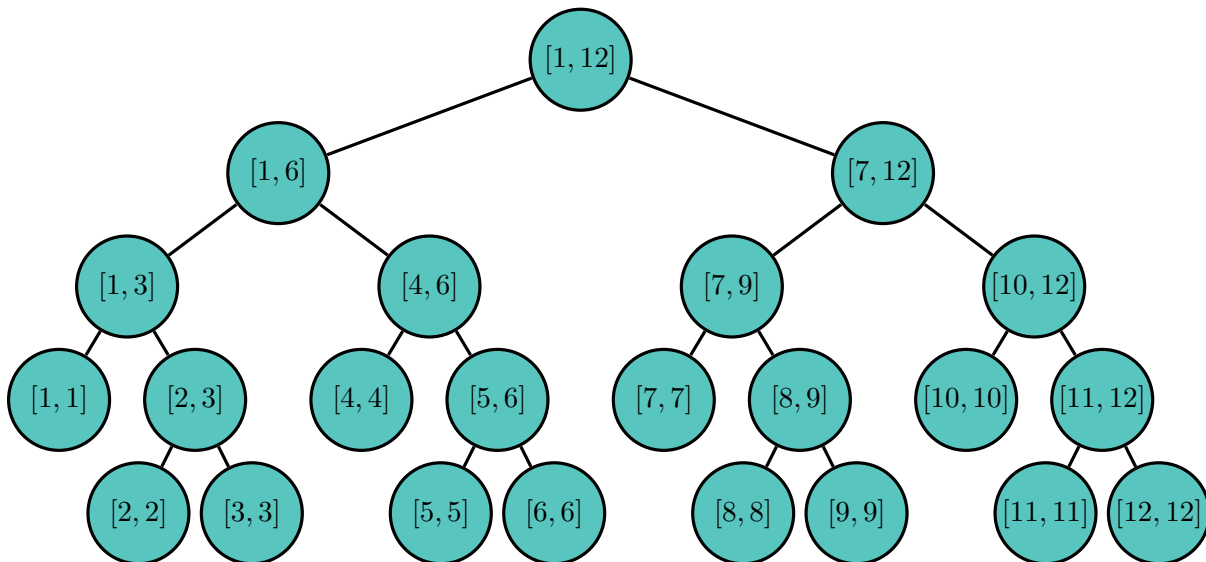
Again, the key idea is if the the interval associated with a node is completely contained within the interval we are querying, we simply return the sum associated with that node. Otherwise, we recurse on the two children. This process is $O(\log n)$ because each level in the tree can have at most two highlighted nodes.

If we wanted to change the third element to 2, we would have to update the highlighted nodes in the following diagram. This process is more straightforward than the querying.



Updating is also $O(\log n)$ as we need to change the values of at most one node in each level in the tree.

I cheated with my example by using a nice power of two, $n = 16$, as the number of elements in the sequence. Of course, the size is not always this nice. One solution is to simply pad the back of the sequence with enough 0s and pretend that the number of elements in the sequence is actually a power of two. For the segment tree, this is not necessary – if a vertex is associated with the range $[a, b]$, we can simply split this range into two, $[a, \lfloor \frac{a+b-1}{2} \rfloor]$ and $[\lfloor \frac{a+b-1}{2} \rfloor + 1, b]$. We then recursively apply this pattern except for when the lower bound of the range is equal to the upper bound. If $n = 12$, the resulting tree would have the following structure.



For this reason, while I used the concept of “building up” on top of our array to introduce the segment tree, any operation we implement will start at the root and recursively trickle down the tree. We see that the segment tree structure does not have to resemble a complete tree at all.

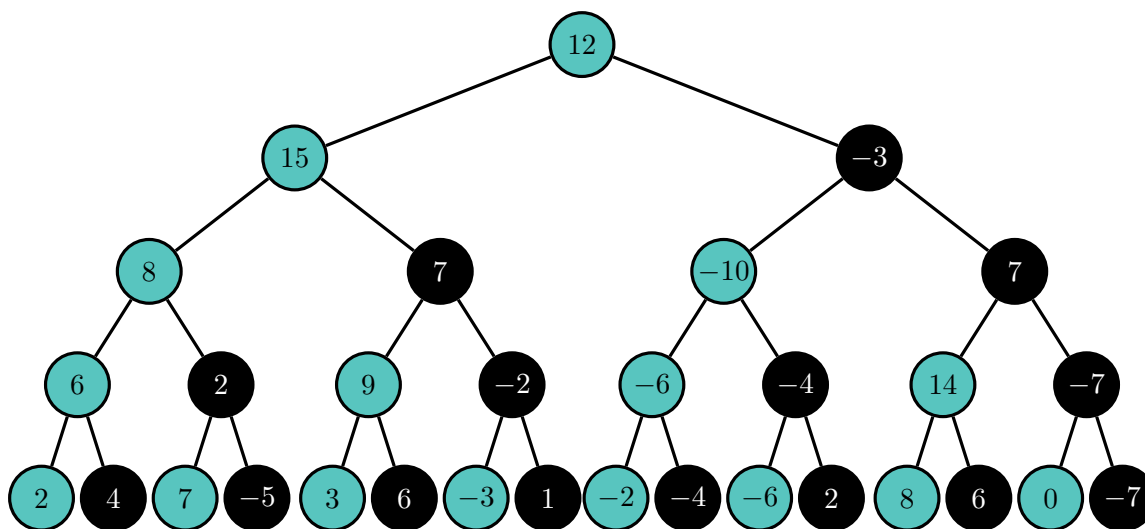
However, with this approach, it is still quite balanced, so we can store a segment tree within an array as we would a heap.

7.2.1 Fenwick Tree

A *Fenwick tree*, or *binary indexed tree (BIT)*, is simply a faster and easier to code segment tree when the operator in question has an inverse. Unfortunately, it's not at all intuitive, so bear with me at first and let the magic of the Fenwick tree reveal itself later. In fact, it is so magical that Richard Peng hates it because it is too gimmicky. The key idea is to compress the data stored within a segment tree in a crazy way that ends up having a really slick implementation using some bit operation tricks.

As discussed earlier, the $+$ operator has an inverse, $-$. Therefore, there is an inherent redundancy, for example, in keeping track of the sum of the first $\frac{n}{2}$ elements, the sum of all n elements, and the sum of the last $\frac{n}{2}$ elements, as we do in the segment tree. If we are given only $\sum_{k=1}^{n/2} a_k$ and $\sum_{k=1}^n a_k$, we can find $\sum_{k=n/2+1}^n a_k$ easily using subtraction.

With this in mind, let's ignore every right child in the tree. We'll mark them as black in the diagram. After that, we'll write out the tree nodes in postfix traversal order, without writing anything whenever we encounter a black node.



2	4	7	-5	3	6	-3	1	-2	-4	-6	2	8	6	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

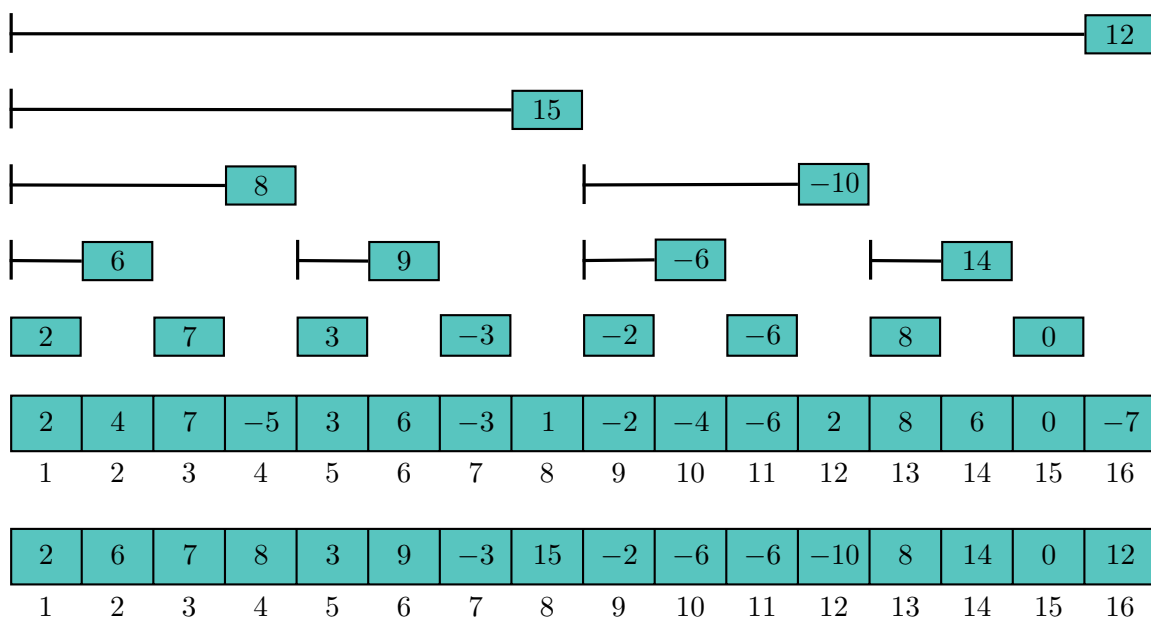
2	6	7	8	3	9	-3	15	-2	-6	-6	-10	8	14	0	12
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Our Fenwick tree is simply this last array. This should be quite confusing – it is not at all clear why this array resembles a tree, and the numbers in the array make no sense whatsoever right now.

Notice that the final position of every unblackened node is just the rightmost black child in its subtree. This leads to the fact that the i th element in the Fenwick tree array is the sum

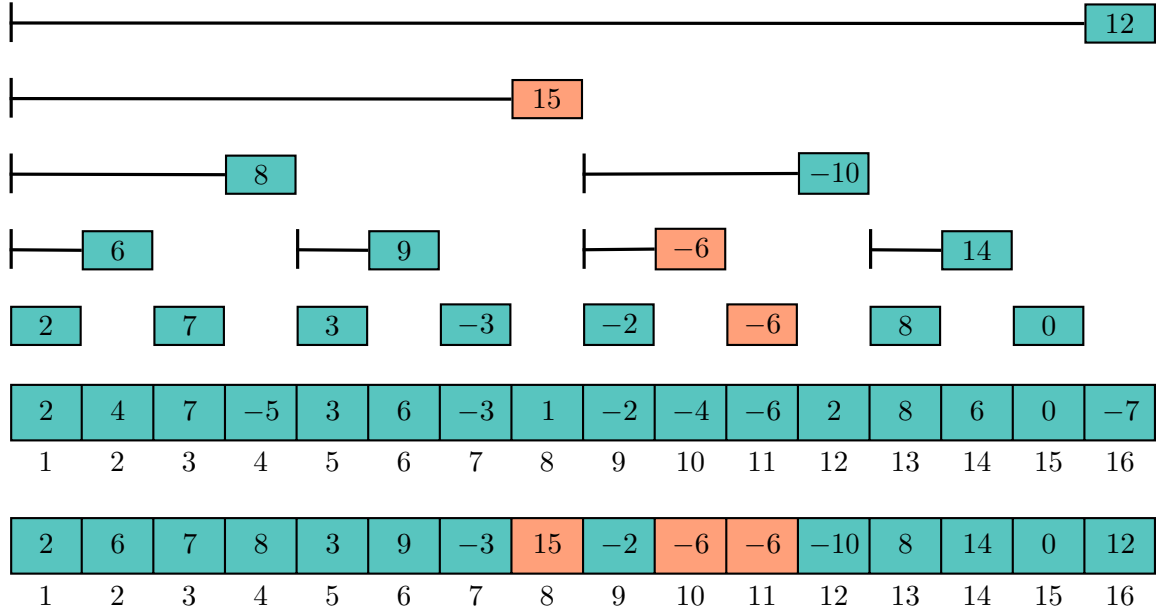
$$b_k = \sum_{k=i-2^{v_2(i)}+1}^i a_k,$$

where $2^{v_2(i)}$ is simply the greatest power of 2 that divides i . Let's look at a new diagram that hopefully will better illustrate this key property of the random array we just came up with.



All the framework is now in place. Now we need to find out how to query and update the Fenwick tree.

Suppose we wanted to find the sum $\sum_{k=1}^{11} a_k$. Let's take a look at the diagram to see which elements we need.

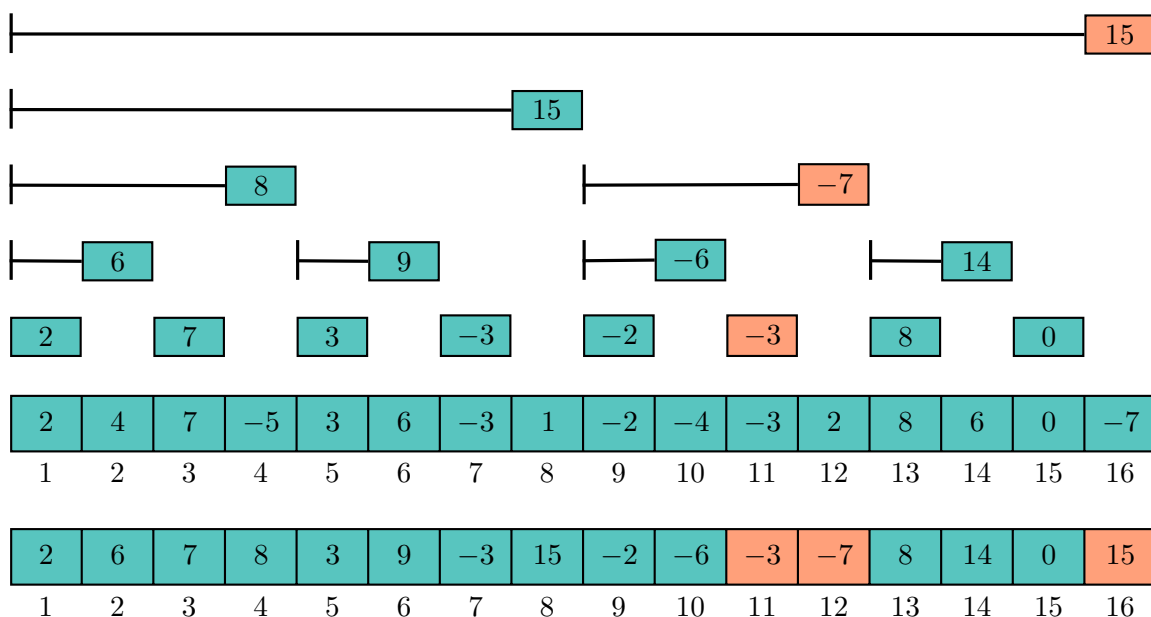


We see that the sum $\sum_{k=1}^{11} a_k = b_8 + b_{10} + b_{11}$. If we look at 11 in binary, we have $11 = 01011_2$. Let's see if we can find a pattern in these numbers in binary:

$$\begin{aligned}
 11 &= 01011_2, \\
 10 &= 11 - 2^{v_2(11)} = 01010_2, \\
 8 &= 10 - 2^{v_2(10)} = 01000_2, \\
 0 &= 8 - 2^{v_2(8)} = 00000_2.
 \end{aligned}$$

So, we can simply subtract $11 - 2^{v_2(11)} = 10 = 01010_2$, find the sum of the first 10 elements, and add b_{11} to that sum to get the sum of the first 11 elements. We see that repeating this process takes off the last 1 in the binary representation of the number i , and since there are at most $\log n + 1$ 1s in the binary representation $\forall i \in [1, n]$, the query operation is $O(\log n)$.

And now for the update operation. Suppose we want to change the value of a_{11} from -6 to -3 . Which numbers will we have to change?



We needed to increment the highlighted values, b_{11} , b_{12} , and b_{16} , by 3. Once again we'll look at 11, 12, and 16 in base 2.

$$11 = 01011_2,$$

$$12 = 01100_2 = 11 + 2^{v_2(11)},$$

$$16 = 10000_2 = 12 + 2^{v_2(12)}.$$

It appears that instead of subtracting the largest dividing power of 2, we are adding. Once again this is an $O(\log n)$ operation.

The real magic in the Fenwick tree is how quickly it can be coded. The only tricky part is finding exactly what $2^{v_2(i)}$ is. But it turns out, by the way bits are arranged in negative numbers, this is just $i \& -i$. With this in mind, here's all the code that's necessary to code a Fenwick tree.

```

1 int[] b = new int[MAXN]; // Fenwick tree stored as array
2 void update(int i, int x) {
3     for( ; i < MAXN; i += i & -i)
4         b[i] += x;
5 }
6 int prefixSum(int i) {
7     int sum = 0;
8     for( ; i > 0; i -= i & -i)
9         sum += b[i];
10    return sum;
11 }
12 int query(int i, int j) {
13     return prefixSum(j) - prefixSum(i - 1);
14 }

```

7.2.2 Lazy Propagation

It is often the case that in addition to performing range queries, we need to be able to perform *range updates*. (Before, we only had to implement point updates.) One extension of our sum problem would require the following two functions:

- $update(i, j, x)$ – increment the value of a_k by x for all $k \in [i, j]$
- $query(i, j)$ – return $\sum_{k=i}^j a_k$.

Some Motivation: \sqrt{n} Blocking

Let's go back to our \sqrt{n} blocking solution and see what changes we can make, and hopefully we can extend this idea back to our segment tree. If we're looking for an $O(\sqrt{n})$ implementation for *update*, we clearly can't perform point updates for all values in the range. The way we sped up *query* was by keeping track of an extra set of data, the sum of all the elements in a bucket, which we used when *the entire bucket was in the query range*.

2	4	7	-5	3	6	-3	1	-2	-4	-6	2	8	6	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

8	7	-10	7
[1, 4]	[5, 8]	[9, 12]	[13, 16]

What can we do with this idea for *update*? Let's see what we can do if an entire bucket were included in the update range. Again, we don't want to touch the original array a at all since that makes the operation linear. Let's try storing some information separately. This other information we're storing is the key idea behind lazy propagation.

With this in mind, highlighted are the elements we'll need for $update(4, 14, 3)$.

2	4	7	-2	3	6	-3	1	-2	-4	-6	2	11	9	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

11	19	2	13
[1, 4]	[5, 8]	[9, 12]	[13, 16]

0	3	3	0
[1, 4]	[5, 8]	[9, 12]	[13, 16]

Note that the sums associated with each bucket must adjust appropriately. In the example, there are four elements per bucket, so when an entire bucket needs to be incremented by 3, a single bucket can increment its sum by $4 \cdot 3 = 12$ easily.

To reiterate, we not storing the actual values of the elements where they were stored in solving the original formulation of the problem. Despite this fact, we are able to calculate what any single value is supposed to be. a_i is simply equal to the $\left\lceil \frac{i}{\sqrt{n}} \right\rceil$ th value stored in the newest third array added to the i th value stored in the first array. Because of this, querying a range works in almost exactly the same way as it did in the original formulation.

Highlighted are the values necessary for $query(7, 15)$.

2	4	7	-2	3	6	-3	1	-2	-4	-6	2	11	9	0	-7
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

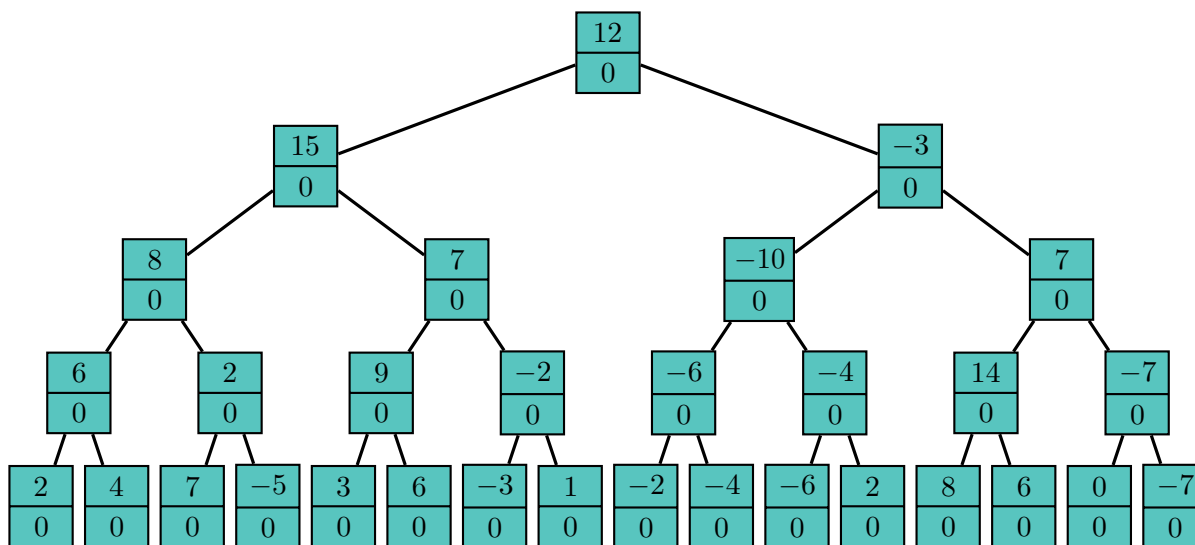
11	19	2	13
[1, 4]	[5, 8]	[9, 12]	[13, 16]

0	3	3	0
[1, 4]	[5, 8]	[9, 12]	[13, 16]

Thus we have achieved $O(\sqrt{n})$ for both range updates and and range queries.

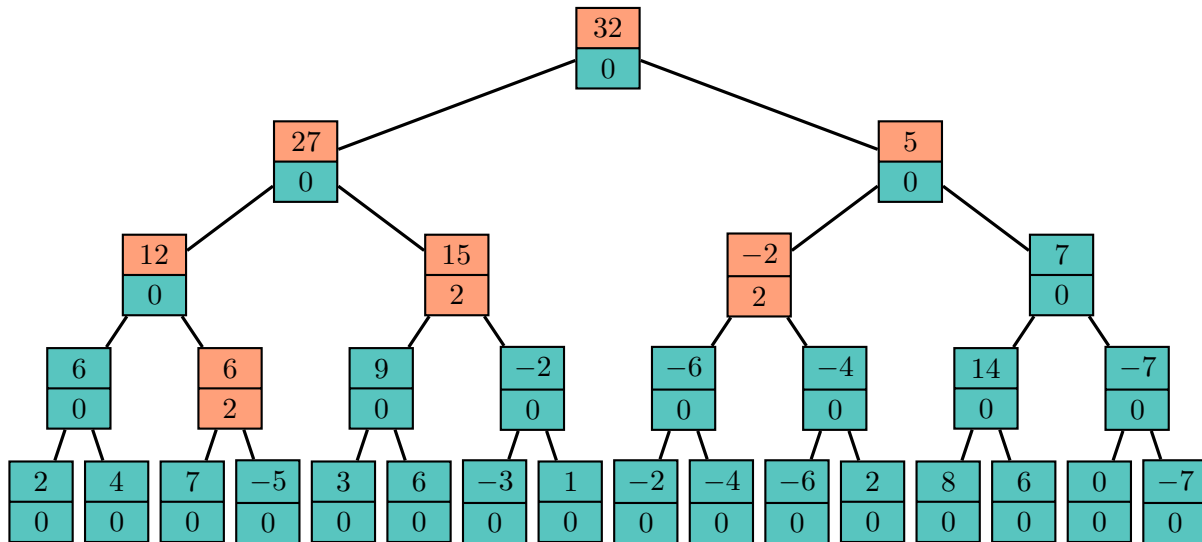
Lazy Propagation on a Segment Tree

Motivated by how we fixed our $O(\sqrt{n})$ solution, let's try adding a similar extra piece of information to our segment tree to try to get an $O(\log n)$ solution. Let's call this extra number the “lazy” number.



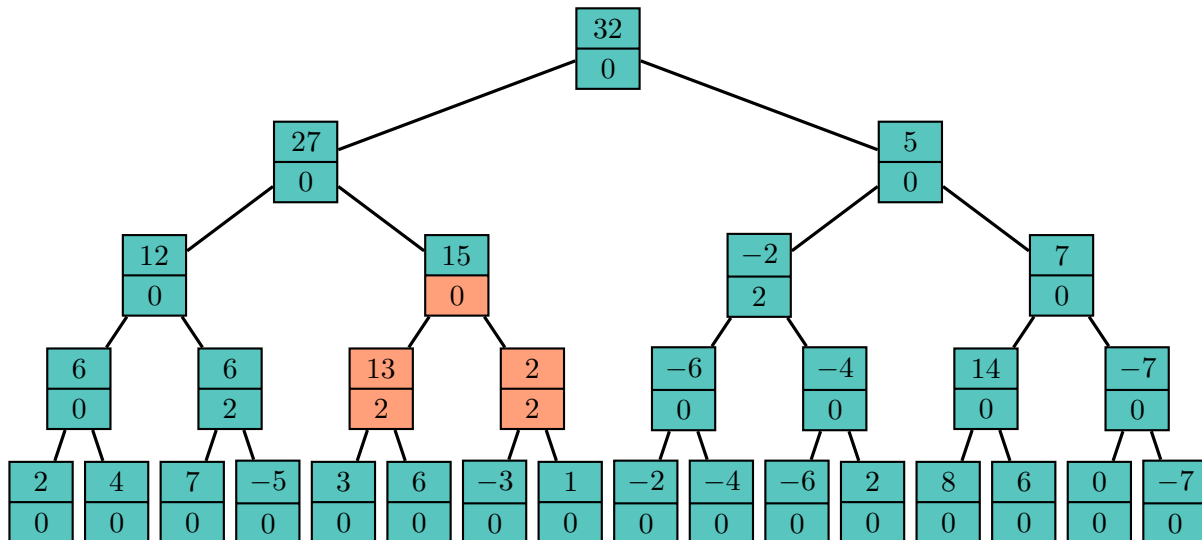
Once again, if the entire range associated with a node is contained within the update interval, we'll just make a note of it on that particular node and not update any of its children. We'll call such a node “lazy.”

Here's the status of the tree after $update(3, 12, 2)$.



When a node is lazy, it indicates that the sum numbers of every other node in its subtree is no longer accurate. In particular, if a node is lazy, the sum number it keeps track of is not equal to the sum of the sum numbers of its children. This means whenever we need to access any node in that subtree, we'll need to update them then. Numbers in the tree therefore can change even after a query. Let's perform a query to illustrate that point.

$query(7, 13)$ requires access to the nodes for the ranges $[7, 8]$, $[9, 12]$, and $[13, 13]$. The nodes for $[9, 12]$ and $[13, 13]$ are up to date and store the correct sum, but the node for $[7, 8]$ does not, as it is a descendent of the node for $[5, 8]$, which has a nonzero lazy number. Highlighted are the nodes we'll need to update as we perform the query. Notice how $[5, 8]$ simply passes its lazy number onto its children, and they update themselves as necessary.



Now, the nodes that we need are all up to date, and we can perform our query. This process is necessary whenever we are performing an operation on an interval that intersects but does not contain the range associated with a lazy node. Since this can only happen for two such nodes (one at either end of the interval in question), both updating and querying change values of at most four nodes per level, so both operations are $O(\log n)$.

7.2.3 Higher Dimensions

7.2.4 Persistent Segment Tree

1. (SPOJ, CPAIR) Given a sequence A_k of N ($1 \leq N \leq 100,000$) non-negative integers ($\forall k, A_k < 1000$), Answer Q ($1 \leq Q \leq 100,000$) queries, where each query consists of three integers, v, a, b . The answer is the number of pairs of integers i and j such that $1 \leq i \leq j \leq N$, $a \leq j - i + 1 \leq b$, and $A_k \geq v$ for every integer k between i , and j , inclusive.

7.3 Queue with Minimum Query

Suppose we wanted a list data structure with the following three operations:

- $add(x)$ – add x to the end of the list.
- $remove()$ – remove the first element in the list.
- $min()$ – return the minimum element in the list.

This is different from a heap since $remove$ does not remove the minimum element. It's pretty easy to find a $O(\log n)$ solution using the data structures we already know. However, it is possible to build a data structure that can do any of these operations with complexity $O(1)$.

To solve this problem, we'll first solve an easier problem. Suppose instead of removing the first element of the list, we had remove the last element; in other words, we needed to build a stack with minimum query instead of a queue. This is a simple task; we'll just use a normal stack, but instead of storing single numbers, we'll store pairs. The pairs will each contain the number we're adding and the minimum element up to that point in the stack.

To build a queue given a stack with minimum query, we'll just have two stacks. When we add an element, we push it to the top of the first stack. When we remove an element, we take it off the top of the second stack. The minimum element in the queue is the smaller element between the minima of either stack.

This seems like it obviously doesn't work – one of the stacks keeps growing, while the other can only shrink. This is not a problem, however; when the second stack runs out of elements, we'll just pop off every element of the first stack and push each onto the second stack. This amounts to one $O(n)$ operation for every n $O(1)$ operations, which averages out to constant time.

7.4 Balanced Binary Search Tree

Recall that a normal binary search tree is not guaranteed to be balanced. Many data structures have been invented to self-balance the binary search tree.

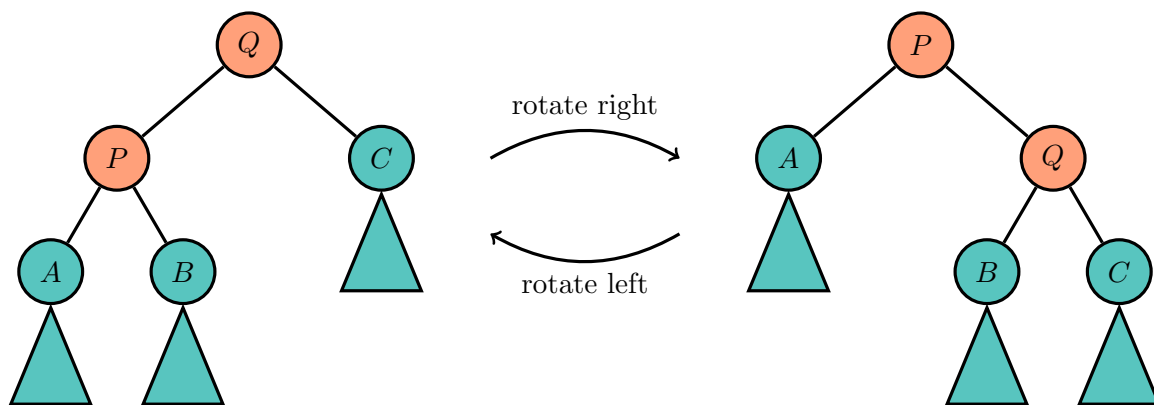
The *red-black tree* was an early development that represents the “just do it” approach, using casework to balance a tree. This results in a simple mathematical proof for the maximum depth of the tree. Unfortunately, coding these up are quite nasty do to the casework involved, so doing so is generally not preferred unless your name is Sreenath Are.

The *splay tree* is a data structure that guarantees amortized logarithmic performance; this means that any single operation is not necessarily linear, but over *any* sequence of length m for $m \geq n$, the performance for all m operations is $O(m \log n)$. Though the proof of the splay tree’s complexity is necessarily more complicated than the red-black tree’s complexity, as it requires amortized analysis, coding the splay tree is much easier than coding the red-black tree. In addition, the splay tree grants us more flexibility than the red-black tree provides, allowing us to build more complicated structures like link-cut trees using splay trees.

The *treap* is a probabilistic data structure that combines the BST with a heap to usually create a balanced tree. Just as quicksort is not guaranteed to run in $O(n \log n)$, so too does the treap not guarantee $O(\log n)$ operations, but will usually result in logarithmic performance.

7.4.1 Tree Rotation

The key idea behind all of these data structures is the use of the *tree rotation*, which is used to change the structure of the tree but will not change the order of the elements (inorder). Maintaining the order of the elements is important, of course, if we wish our tree to remain a binary search tree.

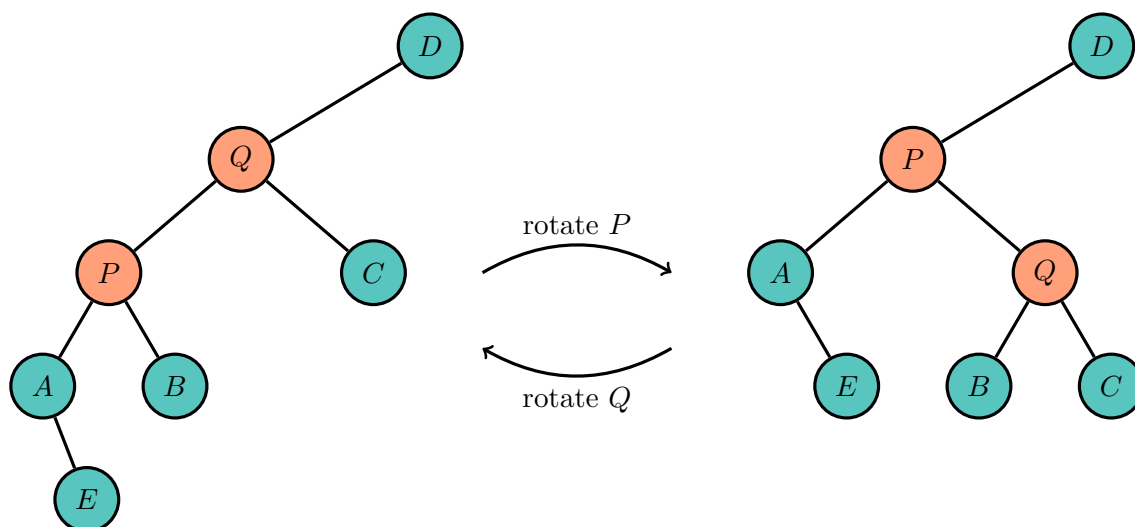


Here, the triangles represent subtrees, as A , B , and C could very well have children of our own, but they are not shown in the diagram. Note that the inorder ordering of the nodes has not been changed.

When we rotate right, we literally take the edge connecting P and Q and rotate it clockwise. Then P becomes the parent of Q where before P was the child of Q . However, this definition of rotation is somewhat cumbersome, as we have different terms for the highly symmetrical rotating right and rotating left. The key characteristic a rotation is we move the lower node up one level.

Thus, I prefer to think of tree rotation as whatever tree rotation, either left or right, will *rotate up* a node. In the diagram, rotating right is analogous to rotating P up, and rotating left is analogous to rotating Q up. Rotating a node up will change the tree such that its former parent is now its child.

The other notable change in the tree structure is the subtree associated with B passes between P and Q upon tree rotation. Finally, tree rotation can happen at any place in the tree, not just at the root. When we rotate P up, we must update the parent of Q to change its child to P .



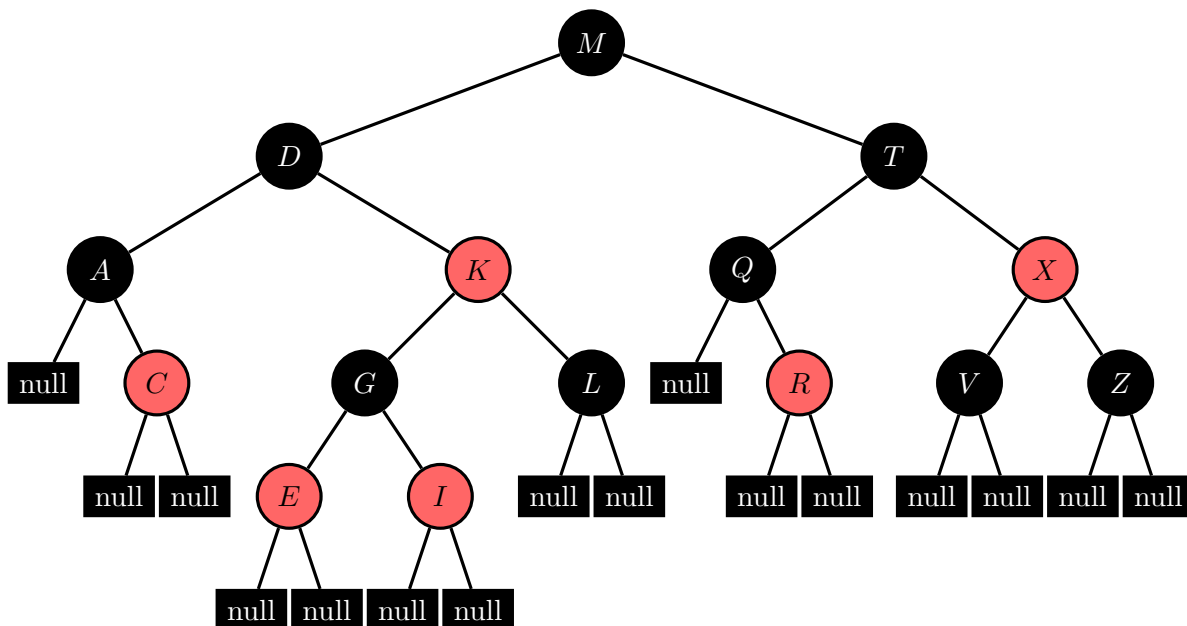
Note that in this example, rotating up P decreases the total height of the tree. We want to somehow systematically rotate up nodes to accomplish this. The following data structures provide such a system.

7.4.2 Red-Black Tree

As stated earlier, the red-black tree represents the “just do it” approach. We want to somehow bound the maximum length from the root to any leaf node by applying a constraint on the tree. In this case, our constraint is a coloring of the nodes (red or black) such certain properties are satisfied. For the red-black tree, the only nodes storing data are non-leaf nodes, and all leaf nodes store “null.” The number of null nodes, then, is $n + 1$, where n is the number of non-null nodes. We do this so that any node storing useful information has two children.

We want our tree to satisfy the following properties:

1. The root is black.
2. All leaf nodes (null) are black.
3. Every red node has two black children. Consequently, a red node has a black parent.
4. Any path from the root to a null node contains the same number of black elements.



Note that every path from the root to a null node contains four black nodes.

The proof of $O(\log n)$ search follows immediately. The shortest possible path contains only black nodes, and the longest possible path contains black and red nodes alternating. Since the number of black nodes in both must be the same, any path is at most twice as long as any other path. As the number of nodes in the tree is $2n + 1$, the number of black nodes m in any path is then bounded below by $2^{2m} - 1 \geq 2n + 1$ and above by $2^m - 1 \leq 2n + 1$. Thus the height of the tree is on the order $O(\log n)$, and we are done.

Thus if our tree maintains its red-black coloring and satisfies the necessary properties, we can guarantee that our tree is balanced. We then consider the two ways we change the state of the tree, insertion and deletion. We can insert and delete in the normal way, but we might need to make changes to the tree after that to restore the red-black properties. We do this through a small number of color flips and tree rotations, which we can handle through casework.

Let's handle insertion first. When we insert a node, it takes the place of a black null leaf node. To maintain property 4, we must color the new node red, as it has two black null children. However, we may have violated some of the other properties, specifically 1 and 3.

We'll call the new node N , its parent P , its uncle (parent's sibling) U , and its grandparent G , if they exist.

We consider the following five cases.

1. N is the root node. That is, N is the first node added to the tree.

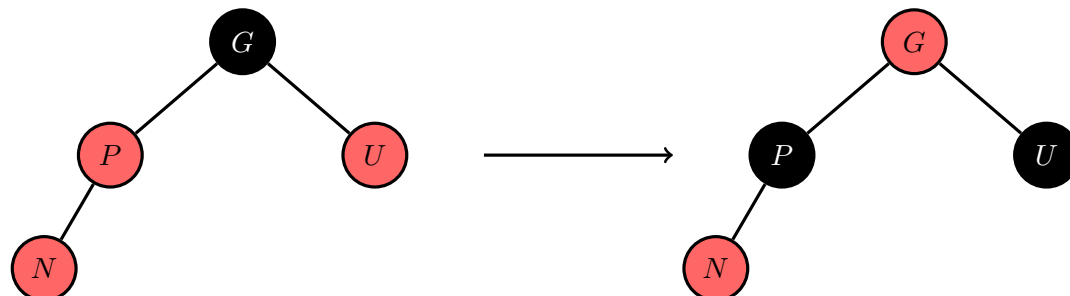
It is easy to just change the color of N to red to restore property 1, and we are done.

2. P is black.

Then property 3 is not violated, and we are done.

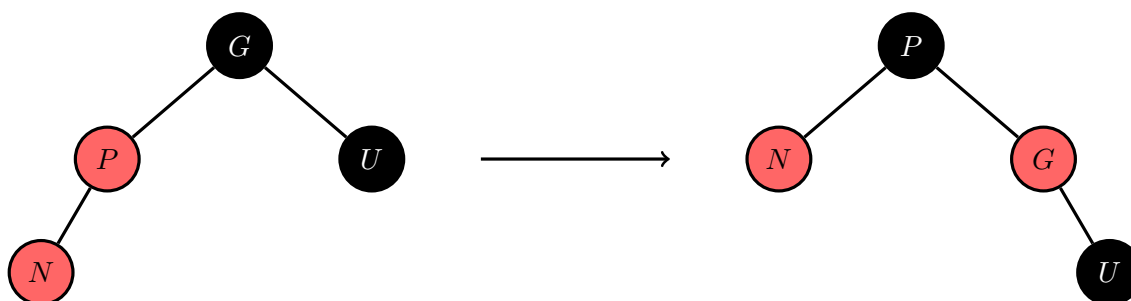
3. P is red, and U is red (G and U exist since P cannot be the root, as the root is black).

As P and U are red, G is black. We simply color P and U black and G red to restore property 3.



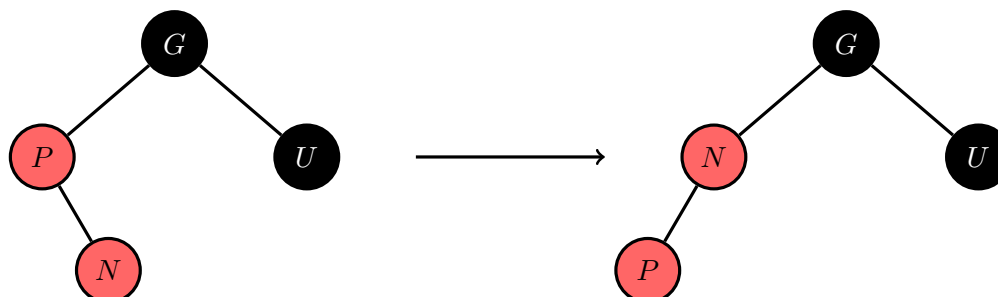
4. P is red, U is black, and N is on the same side of P as P is of G .

G must be black. We rotate up P and color P black and G red.



5. P is red, U is black, and N is on the opposite side of P as P is of G .

G must be black. We rotate up N , and this reduces to case 4.



Thus after we insert, we can always restructure the tree using a constant number of operations to restore the necessary red-black tree properties.

Now we need to work on deletion. Recall how deletion works on a normal binary search tree. If the node we need to replace has two children, we swap the node with the least element in its right subtree, which does not have two children. We then are able to remove that node more easily.

We can do the same thing with the red-black tree. If a node has two non-null children, we swap its value with the least non-null node in its right subtree and remove that node. Thus we reduce the deletion problem to the case where the node we need to remove has at least one null child.

Two cases are very easy.

1. The node we wish to remove is red.

Then the node must have two null leaf nodes as children. Then we remove the node by replacing it and its children with a single null node.

2. The node is black and has a red child.

We replace the node with its child and paint its child red.

The remaining case is the node black with two black null children. We'll first replace the node with a null node N . Then, all paths passing through N are one black node short compared to all other paths.

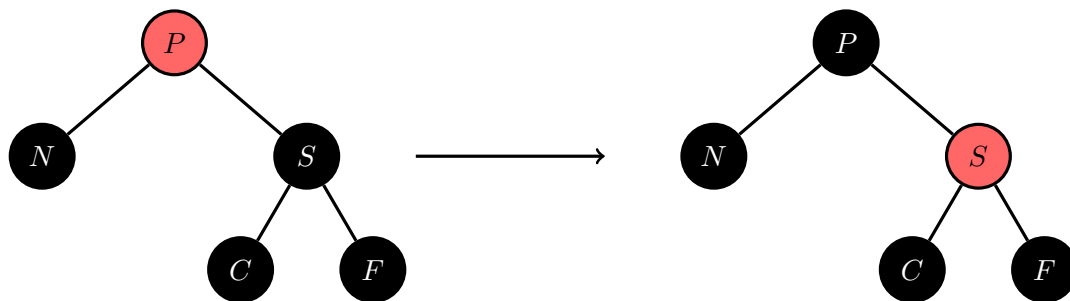
We denote the parent of N as P , its sibling S , and its sibling's children C and F , such that C is on the same side of S as N is of P , if they exist. That is, C is the “closer nephew” child, while F is the farther. We now describe a six-case balancing procedure on the black node N that fixes our problem.

1. N is the root.

We are done, as every path possible must pass through N , so all paths are balanced.

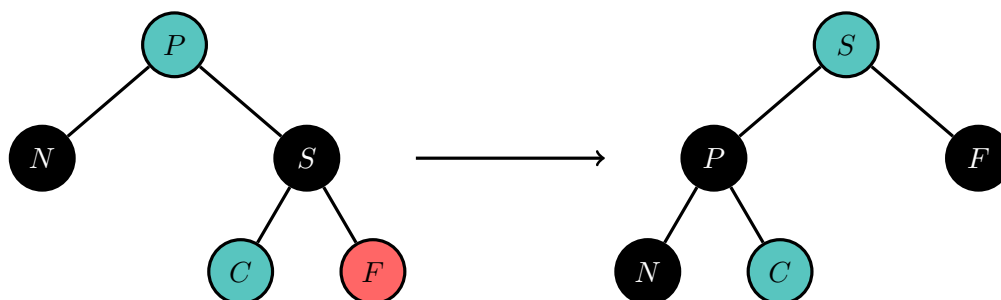
2. P is red and S , C , and F are black.

Then we simply trade the colors of P and S . The number of black nodes for paths passing through S stays the same, but the number of black nodes for paths passing through N increases, so we are done.



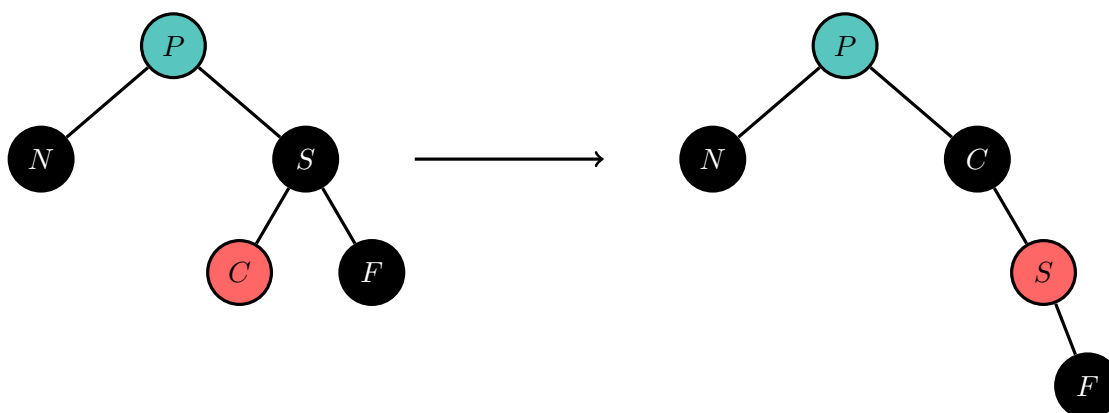
3. S is black and F is red.

Then we color F black and rotate up S , before swapping the colors of P and S . Then paths passing through N gain a black vertex, while paths passing through C and F stay the same.



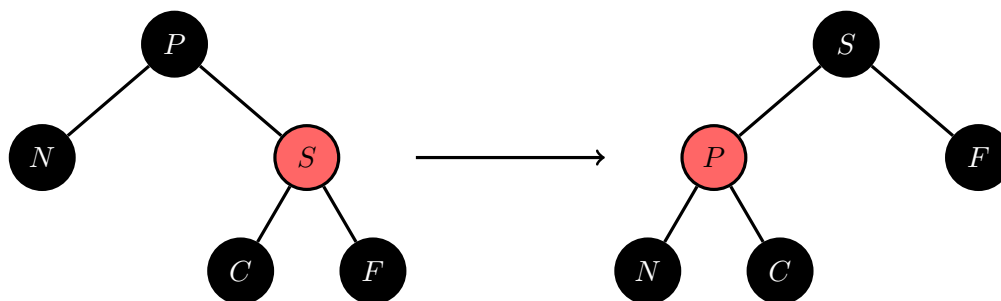
4. S is black, F is black, and C is red.

Then we rotate up C and swap the colors of S and C . Then this reduces to case 3.



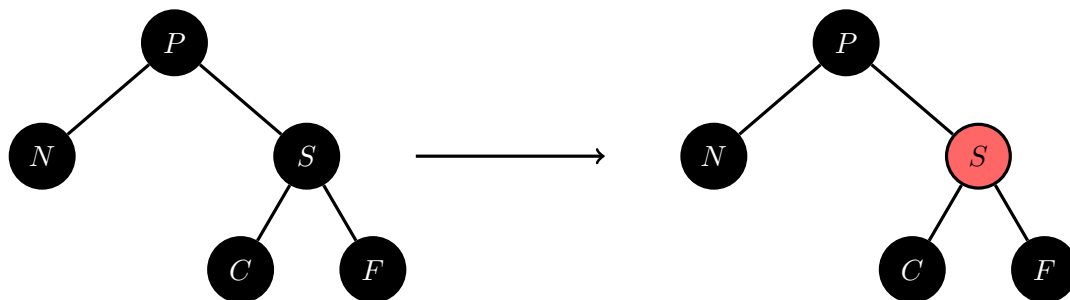
5. S is red.

Then P , C , and F must be black. We rotate S up and swap the colors of P and S . C , the new sibling of N is black, so this then reduces to one of cases 2, 3, or 4.



6. P , S , C , and F are all black.

Then we recolor S red. Now, all paths through P have the same number of black nodes, but each path through P has one less black node than each path not passing through P . Since P is black, and this procedure did not require that N be a leaf node, we can repeat this entire balancing procedure on P .



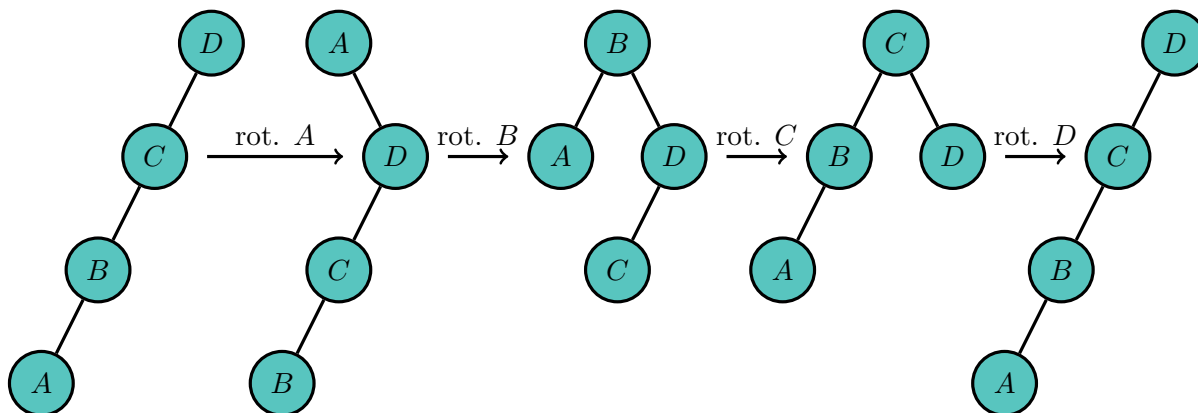
Unlike the balancing for insertions, the balancing for deletions has the potential to call itself on the parent node P . However, balancing following a deletion is still $O(\log n)$ worst case. It is possible to prove that it is $O(1)$ amortized. Regardless, we now have a deletion algorithm that runs in $O(\log n)$ time.

Thus the red-black tree supports standard binary search tree operations, all in $O(\log n)$.

7.4.3 Splay Tree

Remember that rotating a node brings it one level closer to the root. The idea behind a splay tree is to apply a sequence of rotations that rotate a node all the way up to the root. The intuition for doing this is once we access a node, it is easy to perform other operations on it since it will simply be the root of the tree.

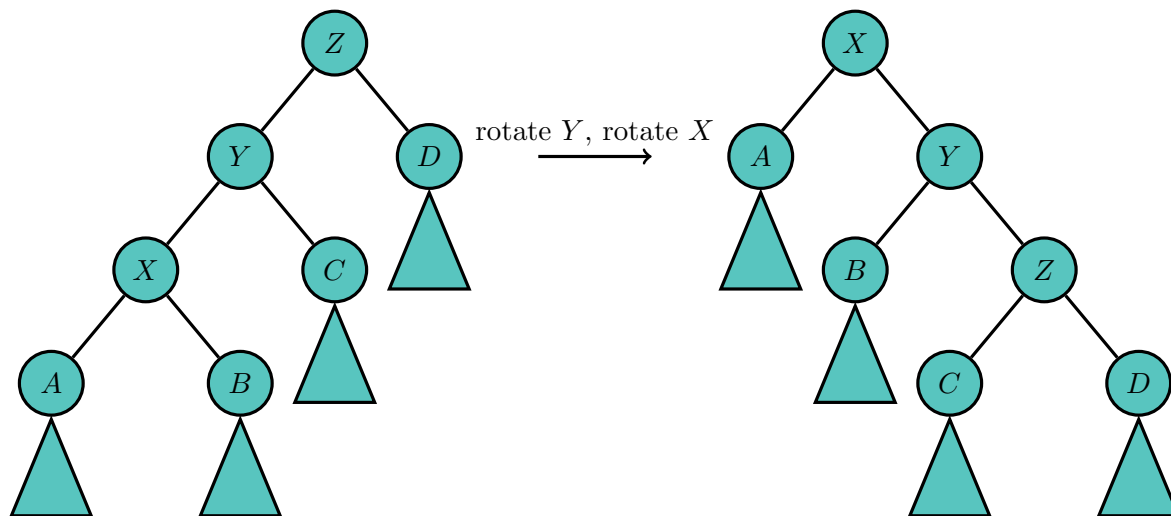
It is not at all clear that rotating nodes to the root will somehow improve the average performance of the tree. That's because rotating the normal way doesn't! To improve the performance of the tree, we expect that our balancing procedure tends to decrease the height of the tree. Unfortunately, simply repeatedly rotating a node up the tree doesn't exactly do that task. Consider a tree that looks like a linked list, and rotating the nodes in the order A, B, C, D , as shown in the diagram. The height of the tree remains linear in magnitude, as at its smallest stage the tree is around $\frac{n}{2}$ in height, so accessing elements remains $O(n)$ on average.



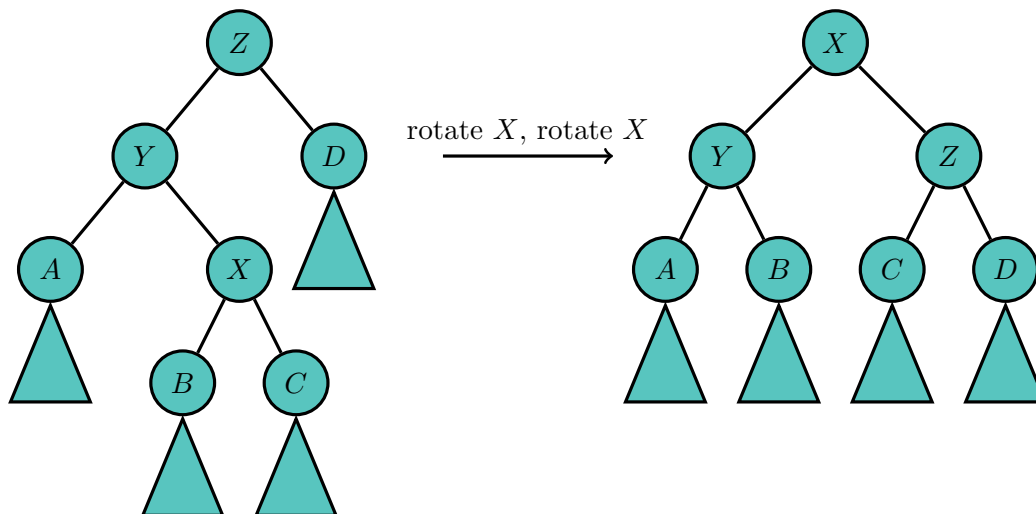
As discussed earlier, simply repeatedly applying the standard rotation clearly is not guaranteed to reduce the height of the tree on average. However, if we simply make one small change, magic happens. The trick of the splay tree is to rotate a node to the root in such a way that the tree has a

tendency to decrease in height. We'll use two compound rotations, and depending on the structure of the tree, we'll use a different one accordingly, to rotate a node to the root.

When a node is a left child and its parent is a left child, or the node is a right child and its parent is a right child, we first rotate up the parent, and then rotate up the node. This sequence of rotations is the only difference between splaying and rotating a node to the root using standard rotation.

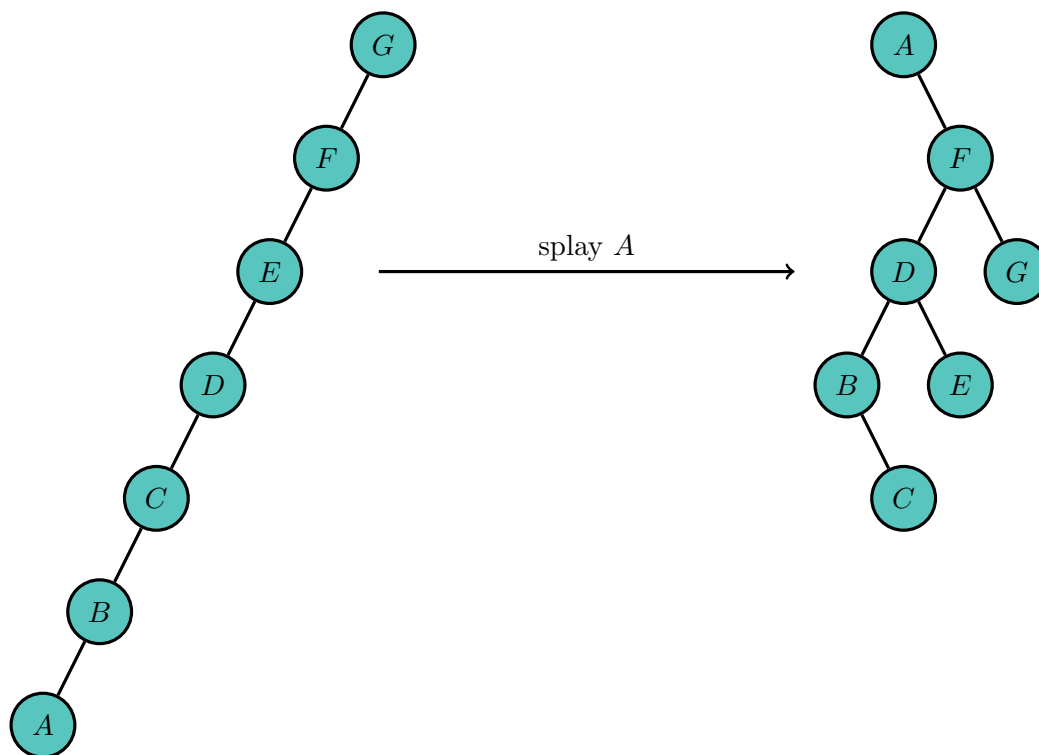


When a node is a left child and its parent is a right child, or the node is a right child and its parent is a left child, we rotate the node up twice, the normal way.



Finally, when a node's parent is the root, so it has no grandparent, we simply rotate up the normal way.

Rotating a node to the root in this way is called *splaying* the node. Thus, splaying a node brings it to the root of the tree. Splaying seems too simple to lead to an amortized $O(\log n)$ solution. To get a better idea of how splaying works, let's see what it does to a tree that looks like a linked list.



Splaying here cuts the height of the tree in half – a huge improvement. Splaying is not guaranteed to decrease the height of the tree, and it is possible for a sequence of splays to even result in a linked-list-like structure. However, given a number m at least equal to the maximum number of nodes n ever in the tree, *any* sequence of m splays runs in $O(m \log n)$. Splaying is then $O(\log n)$, amortized.¹

Then, whenever we access a node, even in insertion and deletion, we splay that node to the top. This makes access, insert, and delete all $O(\log n)$ amortized.

Since splay trees need not satisfy a restrictive coloring, as red-black trees do, we have the freedom to completely change the structure of the tree at a whim. Recall how tedious the casework was to simply add or remove one node at a time in a red-black tree. Because the splay tree is simply a normal binary search tree that rotates nodes up upon access, we can detach an entire subtree from the tree and not worry about any properties of our tree no longer being satisfied.

For this reason, we define the *split* and *join* functions for splay trees.

Given two trees S and T , such that every element in S is less than every element in T , we join S and T into one tree by splaying the largest element in S , so that the root, which is now the largest element, has no right child. Then, we set its right child to the root of T , resulting in a single tree with elements from both S and T .

Given a tree and a value v , we split the tree in two by splaying the greatest element not greater than v to the root and detaching the right subtree from the rest. This results in two trees, one containing all elements at most v , and the other containing all elements greater than v .

¹If you're curious to see the proof of this, a rather in depth explanation can be found here: <http://www.cs.cmu.edu/afs/cs/academic/class/15859-f05/www/documents/splay-trees.txt>

Both of these operations are $O(\log n)$ amortized. It is possible to implement insert and delete using split and join.

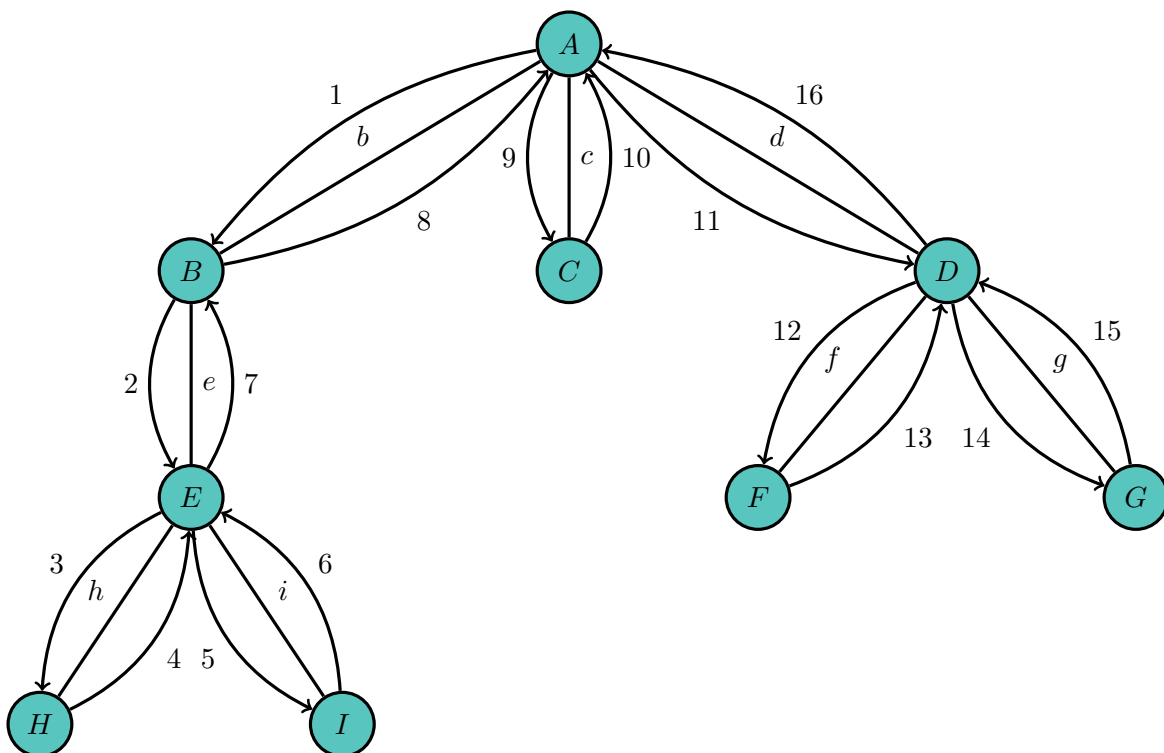
Since a binary search tree stores an ordered set, it is incredibly useful for storing a dynamic list of elements where the length can change and elements need to be added or removed from any point in the list, not just the beginning or the end. Splay trees are incredibly useful because of the split and join operations, as they allow us to remove consecutive elements from the list represented by our tree.

For example, to remove elements indexed in the range $[i, j]$, we split at index i and split at index $j + 1$ to get three splay trees, one representing $[1, i - 1]$, another $[i, j]$, and the last $[j + 1, n]$. We can then merge the first and third trees together, to get two trees, one representing $[i, j]$ and the other representing the original list with $[i, j]$ removed. A similar procedure can insert a list within another list as a contiguous block. Both of these operations can be completed in $O(\log n)$ amortized using a splay tree.

7.4.4 Treap

7.5 Euler Tour Technique

The *Euler tour technique* is a method of representing a tree (not necessarily binary) as a list. The idea is that a tree is uniquely determined by its DFS traversal order, going both down and up the tree. We'll keep track of the order in which we traverse the edges. Since each edge is traversed twice, each edge will appear in our list twice. We'll also name each edge by the child vertex in the parent-child pair.



The edge traversal order is then described by the ordered list below.

b_1	e_1	h_1	h_2	i_1	i_2	e_2	b_2	c_1	c_2	d_1	f_1	f_2	g_1	g_2	d_2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

We see a pattern in this list – the subtree of a node is contained between the two edges representing that node in the list, inclusive. For example, from the first e to the second e , the entire subtree of E is contained within the range $[2, 7]$.

7.5.1 Euler Tour Tree

The *Euler tour tree* uses this idea. Because we see the relationship between the list and the subtrees of the tree, we'll just have the list store the vertex names. To complete this pattern for the subtree of A , which is the entire tree, we'll simply add that letter to the beginning and end of the list.

A_1	B_1	E_1	H_1	H_2	I_1	I_2	E_2	B_2	C_1	C_2	D_1	F_1	F_2	G_1	G_2	D_2	A_2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

Now suppose we wanted to perform some operations on a forest, or collection of trees. The first operation is to *find* the root of the tree containing a vertex v . The second is to *link* the trees of two vertices u, v together by making v the parent of u . The third is to *cut* the subtree of v away from the rest of the tree by deleting the edge from v to its parent.

Note that, without the cut operation, this is simply union-find. However, because of the cut operation, we have to maintain the structure of the original forest, as otherwise we lose track of the actual parent of any given node.

We see that all operations are not simply on a single vertex but on an entire subtree. However, if we look at what each of the operations does to the Euler tour lists representing our forest, find returns the first element in the list, link places one ordered list within another, and cut removes a large contiguous section of one ordered list.

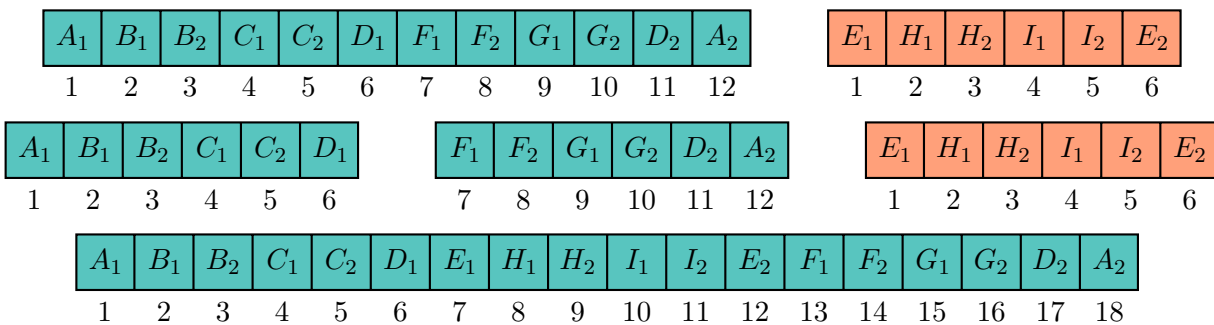
To cut E from the tree, we need to split the list in two:

A_1	B_1	E_1	H_1	H_2	I_1	I_2	E_2	B_2	C_1	C_2	D_1	F_1	F_2	G_1	G_2	D_2	A_2
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

A_1	B_1		E_1	H_1	H_2	I_1	I_2	E_2		B_2	C_1	C_2	D_1	F_1	F_2	G_1	G_2	D_2	A_2
1	2		3	4	5	6	7	8		9	10	11	12	13	14	15	16	17	18

A_1	B_1	B_2	C_1	C_2	D_1	F_1	F_2	G_1	G_2	D_2	A_2		E_1	H_1	H_2	I_1	I_2	E_2
1	2	3	4	5	6	7	8	9	10	11	12		1	2	3	4	5	6

Let's see what happens when we link E to D . We need to split the first list immediately after D_1 .



We just learned about a data structure that maintains a set of ordered elements that can split and merge quickly. Splay trees can maintain an ordered list, and the split and join operations on splay trees can easily implement link and cut. Each tree of our forest is then represented by a splay tree maintaining that tree's Euler tour list, and we can support each of the three necessary operations in $O(\log n)$ amortized time.

7.6 2^n Jump Pointers

7.7 Fractional Cascading

7.8 Heavy-Light Decomposition

Brute-force algorithms on trees are often fast when the tree is nearly complete, or has small depth. Unfortunately, brute force is far too slow when the tree becomes more and more unbalanced and approximates a linked list.

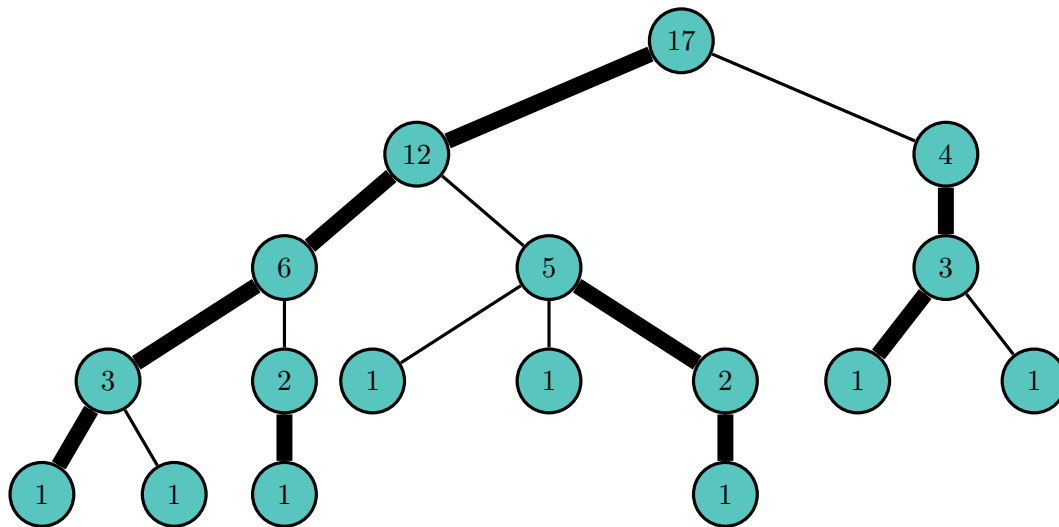
Certain problems can be solved very nicely in a list with segment trees and similar data structures. Heavy-light decomposition provides us with a way to exploit the fact that long “chains,” which are present in unbalanced trees, might be slow in brute force but can be sped up with other data structures.

Heavy-light decomposition provides us a way to dynamically find the lowest common ancestor as well as a way to prove time complexities about tree data structures, like link-cut trees.

Heavy-light decomposition is a coloring of all the edges in a binary tree either heavy or light. For each vertex v , let $s(v)$ denote the number of nodes in the subtree with v as its head. Then, if u, w are the children of v , $s(v) = s(u) + s(w) + 1$.

We see that the smaller child u or w must have less than half the subtree size as v , and so that child exhibits qualities similar to those of a completely balanced tree. We color the edge connecting v to its lesser child *light*. We color the other edge *heavy*.

We see that from any node, the number of light edges needed to reach the head of the tree is at most $\log n$, as each light edge doubles the subtree size. Then, the number of “heavy chains” along the upwards path from any node is also of the order $\log n$.



7.10 Top Tree

7.11 Link-Cut Cactus

Cacti are scary.

Chapter 8

Math

Algorithms here exhibit a different flavor than the graph theory, string, or geometry algorithms from earlier. This chapter is placed towards the end because material here doesn't really fit in any other section or the USACO canon, *not* because this material is particularly difficult.

8.1 Number Theory

In this section we explore ways for computers to quickly compute some useful mathematical quantities.

8.2 Combinatorial Games

https://activities.tjhsst.edu/sct/lectures/1314/impartial_games_12_06_13.pdf

8.3 Karatsuba

8.4 Matrices

8.5 Fast Fourier Transform

Chapter 9

Problems

Problems, to be worked through as you progress through the text.

9.1 Bronze

USACO bronze is quite ad hoc. Knowing basic data structures in the standard library helps.

9.2 Silver

USACO silver tends to have the most standard problems. Silver tests knowledge of basic algorithms and coding ideas. Silver questions also sometimes require knowledge of the language, like familiarity with standard library data structures.

9.2.1 Complete Search

9.2.2 Greedy

9.2.3 Standard Dynamic Programming

1. (USACO Training) There is a long list of stalls, some of which need to be covered with boards. You can use up to N ($1 \leq N \leq 50$) boards, each of which may cover any number of consecutive stalls. Cover all the necessary stalls, while covering as few total stalls as possible.
2. (IOI 1996) You are given a three-valued (1, 2, or 3) sequence of length up to 1000. Find a minimum set of exchanges to put the sequence in sorted order.
3. (Samir Khuller) We are given N jobs, each of which requires one unit of time to complete. The i th job opens at some time t_i and must be completed by some deadline d_i , where $t_i, d_i \in \mathbb{Z}$. Given that only one job can be completed at a time, determine if all N can be completed by their deadlines.

9.2.4 Standard Graph Theory

Flood Fill

Shortest Path

1. (USACO Training Pages, butter) Farmer John owns a collection of pastures with weighted edges between some pairs of locations. Each pasture is inhabited by a cow, and the cows wish to all congregate at one of the pastures. Find the pasture at which the cows should meet in order to minimize combined travel distance.
2. (USACO February 2012, relocate) FJ is moving! He is trying to find the best place to build a new farm so as to minimize his daily travel time.

The region to which FJ plans to move has N towns ($1 \leq N \leq 10,000$). There are M bi-directional roads ($1 \leq M \leq 50,000$) connecting certain pairs of towns. All towns are reachable from each-other via some combination of roads. FJ needs your help selecting the best town as the home for his new farm.

There are markets in K of the towns ($1 \leq K \leq 5$) that FJ wants to visit every day. In particular, every day he plans to leave his new farm, visit the K towns with markets, and then return to his farm. FJ can visit the markets in any order he wishes. When selecting a town in which to build his new farm, FJ wants to choose only from the $N - K$ towns that do not have markets, since housing prices are lower in those towns.

Please help FJ compute the minimum distance he will need to travel during his daily schedule, if he builds his farm in an optimal location and chooses his travel schedule to the markets as smartly as possible.

3. (USACO December 2012, mroute) Farmer John's farm has an outdated network of M pipes ($1 \leq M \leq 500$) for pumping milk from the barn to his milk storage tank. He wants to remove and update most of these over the next year, but he wants to leave exactly one path worth of pipes intact, so that he can still pump milk from the barn to the storage tank.

The pipe network is described by N junction points ($1 \leq N \leq 500$), each of which can serve as the endpoint of a set of pipes. Junction point 1 is the barn, and junction point N is the storage tank. Each of the M bi-directional pipes runs between a pair of junction points, and has an associated latency (the amount of time it takes milk to reach one end of the pipe from the other) and capacity (the amount of milk per unit time that can be pumped through the pipe in steady state). Multiple pipes can connect between the same pair of junction points.

For a path of pipes connecting from the barn to the tank, the latency of the path is the sum of the latencies of the pipes along the path, and the capacity of the path is the minimum of the capacities of the pipes along the path (since this is the "bottleneck" constraining the overall rate at which milk can be pumped through the path). If FJ wants to send a total of X units of milk through a path of pipes with latency L and capacity C , the time this takes is therefore $L + \frac{X}{C}$.

Given the structure of FJ's pipe network, please help him select a single path from the barn to the storage tank that will allow him to pump X units of milk in a minimum amount of total time.

4. (IOI 1999, Traffic Lights) In the city of Dingilville the traffic is arranged in an unusual way. There are junctions and roads connecting the junctions. There is at most one road between any two different junctions. There is no road connecting a junction to itself. Travel time for a road is the same for both directions. At every junction there is a single traffic light that is either blue or purple at any moment. The color of each light alternates periodically: blue for certain duration and then purple for another duration. Traffic is permitted to travel down the road between any two junctions, if and only if the lights at both junctions are the same color at the moment of departing from one junction for the other. If a vehicle arrives at a junction just at the moment the lights switch it must consider the new colors of lights. Vehicles are allowed to wait at the junctions. You are given the city map which shows

- the travel times for all roads (integers),
- the durations of the two colors at each junction (integers)
- the initial color of the light and the remaining time (integer) for this color to change at each junction.

Your task is to find a path which takes the minimum time from a given source junction to a given destination junction for a vehicle when the traffic starts. In case more than one such path exists you are required to report only one of them.

Minimal Spanning Tree

1. (USACO March 2014, irrigation) Due to a lack of rain, Farmer John wants to build an irrigation system to send water between his N fields ($1 \leq N \leq 2000$).

Each field i is described by a distinct point (x_i, y_i) in the 2D plane, with $0 \leq x_i, y_i \leq 1000$. The cost of building a water pipe between two fields i and j is equal to the squared Euclidean distance between them:

$$(x_i - x_j)^2 + (y_i - y_j)^2$$

FJ would like to build a minimum-cost system of pipes so that all of his fields are linked together – so that water in any field can follow a sequence of pipes to reach any other field.

Unfortunately, the contractor who is helping FJ install his irrigation system refuses to install any pipe unless its cost (squared Euclidean length) is at least C ($1 \leq C \leq 1,000,000$).

Please help FJ compute the minimum amount he will need pay to connect all his fields with a network of pipes.

2. (USACO February 2015, superbull) Bessie and her friends are playing hoofball in the annual Superbull championship, and Farmer John is in charge of making the tournament as exciting as possible. A total of N ($1 \leq N \leq 2000$) teams are playing in the Superbull. Each team is assigned a distinct integer team ID in the range $[1, 2^{30} - 1]$ to distinguish it from the other teams. The Superbull is an elimination tournament – after every game, Farmer John chooses which team to eliminate from the Superbull, and the eliminated team can no longer play in any more games. The Superbull ends when only one team remains.

Farmer John notices a very unusual property about the scores in matches! In any game, the combined score of the two teams always ends up being the bitwise exclusive OR (XOR) of

the two team IDs. For example, if teams 12 and 20 were to play, then 24 points would be scored in that game, since $01100XOR10100 = 11000$.

Farmer John believes that the more points are scored in a game, the more exciting the game is. Because of this, he wants to choose a series of games to be played such that the total number of points scored in the Superbull is maximized. Please help Farmer John organize the matches.

3. (SPOJ, INVENT) Given tree with N ($1 \leq N \leq 15,000$) vertices, find the minimum possible weight of a complete graph (a graph where every pair of vertices is connected) such that the given tree is its unique minimum spanning tree.

Union-Find

1. (USACO February 2013, tractor) One of Farmer John's fields is particularly hilly, and he wants to purchase a new tractor to drive around on it. The field is described by an $N \times N$ grid of non-negative integer elevations ($1 \leq N \leq 500$). A tractor capable of moving from one grid cell to an adjacent cell (one step north, east, south, or west) of height difference D costs exactly D units of money.

FJ would like to pay enough for his tractor so that, starting from some grid cell in his field, he can successfully drive the tractor around to visit at least half the grid cells in the field (if the number of total cells in the field is odd, he wants to visit at least half the cells rounded up). Please help him compute the minimum cost necessary for buying a tractor capable of this task.

2. (CF 266E) There are n ($1 \leq n \leq 10^5$) employees working in company "X" (let's number them from 1 to n for convenience). Initially the employees didn't have any relationships among each other. On each of m ($1 \leq m \leq 10^5$) next days one of the following events took place:
 - either employee y became the boss of employee x (at that, employee x didn't have a boss before);
 - or employee x gets a packet of documents and signs them; then he gives the packet to his boss. The boss signs the documents and gives them to his boss and so on (the last person to sign the documents sends them to the archive);
 - or comes a request of type "determine whether employee x signs certain documents".

Your task is to write a program that will, given the events, answer the queries of the described type. At that, it is guaranteed that throughout the whole working time the company didn't have cyclic dependencies.

Euler Tour

1. (USACO Training, Airplane Hopping) Given a collection of cities, along with the flights between those cities, determine if there is a sequence of flights such that you take every flight exactly once, and end up at the place you started. (In other words, find an Eulerian circuit on a directed graph.)

2. (USACO Training, Cows on Parade) Farmer John has two types of cows: black Angus and white Jerseys. While marching 19 of their cows to market the other day, John's wife Farmeress Joanne, noticed that all 16 possibilities of four successive black and white cows (e.g., *bbbb*, *bbbw*, *bbwb*, *bbww*, ..., *www*) were present. Of course, some of the combinations overlapped others.

Given N ($2 \leq N \leq 15$), find the minimum length sequence of cows such that every combination of N successive black and white cows occurs in that sequence.

(The answer is not hard to guess, but use Eulerian circuits to prove that it is correct.)

9.2.5 Easy Computational Geometry

9.3 Gold

USACO gold problems generally fall into two families; the first tests knowledge of more complex data structures not in the standard library, and the second tests cleverness with more nonstandard greedy or dynamic programming strategies.

9.3.1 2^n Dynamic Programming

9.3.2 Binary Search

1. (USACO March 2014, sabotage) Farmer John's arch-nemesis, Farmer Paul, has decided to sabotage Farmer John's milking equipment!

The milking equipment consists of a row of N ($3 \leq N \leq 100,000$) milking machines, where the i th machine produces M_i units of milk ($1 \leq M_i \leq 10,000$). Farmer Paul plans to disconnect a contiguous block of these machines – from the i th machine up to the j th machine ($2 \leq i \leq j \leq N - 1$); note that Farmer Paul does not want to disconnect either the first or the last machine, since this will make his plot too easy to discover. Farmer Paul's goal is to minimize the average milk production of the remaining machines. Farmer Paul plans to remove at least 1 cow, even if it would be better for him to avoid sabotage entirely.

Fortunately, Farmer John has learned of Farmer Paul's evil plot, and he is wondering how bad his milk production will suffer if the plot succeeds. Please help Farmer John figure out the minimum average milk production of the remaining machines if Farmer Paul does succeed.

2. (Matthew Savage) Ashley's journey through Unova is just beginning, and she has just picked her first Pokémon! Unfortunately, not knowing much about them, she picked a Snivy, and a particularly stubborn and unfriendly one at that.

Being Ashley, she decides to try to win over the Snivy in the only way she knows how – baked goods.

Ashley knows r ($0 \leq r \leq 1000$) recipes for Poképufts. Each recipe r_i has a deliciousness rating d_i ($0 \leq d_i \leq 1000$) and requires some combination of the I ($0 \leq I \leq 1000$) available ingredients. (More specifically, the recipe r_i uses the quantity I_{ij} ($0 \leq I_{ij} \leq 1000$) of ingredient I_j .)

Ashley has some amount of each ingredient I_j on hand A_j ($0 \leq A_j \leq 10^9$) and can buy more from the nearby store for a price of c_j ($1 \leq c_j \leq 10^9$) dollars per unit using the M ($0 \leq M \leq 10^{12}$) dollars she currently has.

Of course, Ashley also has limited supplies and therefore can only produce Poképufts from a single recipe. However, she can make as many as she wants, in integer increments.

We define “total deliciousness” (D) to be the sum of the deliciousnesses of the individual Poképufts that Ashley has baked.

Ashley wants to have the best chance possible with Snivy, and therefore would like to know - what is the maximum possible deliciousness ($\max(D)$) that she can produce?

Note: there is a “just do it” solution that is faster than the binary search by a log factor. It is also *much* more annoying to code; so annoying that I was unable to debug my “just do it” solution in the actual contest environment. I included this problem as an exercise to demonstrate how easy binary searching on the answer can be.

3. (CF 287B) Vova, the Ultimate Thule new shaman, wants to build a pipeline. As there are exactly n ($1 \leq n \leq 10^{18}$) houses in Ultimate Thule, Vova wants the city to have exactly n pipes, each such pipe should be connected to the water supply. A pipe can be connected to the water supply if there’s water flowing out of it. Initially Vova has only one pipe with flowing water. Besides, Vova has several splitters.

A splitter is a construction that consists of one input (it can be connected to a water pipe) and x output pipes. When a splitter is connected to a water pipe, water flows from each output pipe. You can assume that the output pipes are ordinary pipes. For example, you can connect water supply to such pipe if there’s water flowing out from it. At most one splitter can be connected to any water pipe.

Vova has one splitter of each kind: with $2, 3, 4, \dots, k$ ($2 \leq k \leq 10^9$) outputs. Help Vova use the minimum number of splitters to build the required pipeline or otherwise state that it’s impossible.

Vova needs the pipeline to have exactly n pipes with flowing out water. Note that some of those pipes can be the output pipes of the splitters.

4. (IOI 2009) Mecho the bear has found a little treasure - the bees’ secret honeypot, which is full of honey! He was happily eating his newfound treasure until suddenly one bee saw him and sounded the bee alarm. He knows that at this very moment hordes of bees will emerge from their hives and start spreading around trying to catch him. He knows he has to leave the honeypot and go home quickly, but the honey is so sweet that Mecho doesn’t want to leave too soon. Help Mecho determine the latest possible moment when he can leave.

Mecho’s forest is represented by a square grid of N by N ($1 \leq N \leq 800$) unit cells, whose sides are parallel to the north-south and east-west directions. Each cell is occupied by a tree, by a patch of grass, by a hive or by Mecho’s home. Two cells are considered adjacent if one of them is immediately to the north, south, east or west of the other (but not on a diagonal). Mecho is a clumsy bear, so every time he makes a step, it has to be to an adjacent cell. Mecho can only walk on grass and cannot go through trees or hives, and he can make at most S ($1 \leq S \leq 1000$) steps per minute. At the moment when the bee alarm is sounded, Mecho is in the grassy cell containing the honeypot, and the bees are in every cell containing a hive (there

may be more than one hive in the forest). During each minute from this time onwards, the following events happen in the following order:

- If Mecho is still eating honey, he decides whether to keep eating or to leave. If he continues eating, he does not move for the whole minute. Otherwise, he leaves immediately and takes up to S steps through the forest as described above. Mecho cannot take any of the honey with him, so once he has moved he cannot eat honey again.
- After Mecho is done eating or moving for the whole minute, the bees spread one unit further across the grid, moving only into the grassy cells. Specifically, the swarm of bees spreads into every grassy cell that is adjacent to any cell already containing bees. Furthermore, once a cell contains bees it will always contain bees (that is, the swarm does not move, but it grows).

In other words, the bees spread as follows: When the bee alarm is sounded, the bees only occupy the cells where the hives are located. At the end of the first minute, they occupy all grassy cells adjacent to hives (and still the hives themselves). At the end of the second minute, they additionally occupy all grassy cells adjacent to grassy cells adjacent to hives, and so on. Given enough time, the bees will end up simultaneously occupying all grassy cells in the forest that are within their reach.

Neither Mecho nor the bees can go outside the forest. Also, note that according to the rules above, Mecho will always eat honey for an integer number of minutes.

The bees catch Mecho if at any point in time Mecho finds himself in a cell occupied by bees.

Write a program that, given a map of the forest, determines the largest number of minutes that Mecho can continue eating honey at his initial location, while still being able to get to his home before any of the bees catch him.

9.3.3 Segment Tree

9.3.4 More Standard Graph Theory

Maximum Flow and Variants

Strongly Connected Components

Other Graph Theory

1. (USACO Open 2008, nabor) Farmer John has N ($1 \leq N \leq 100,000$) cows who group themselves into “Cow Neighborhoods”. Each cow is at a unique rectilinear coordinate, on a pasture whose x and y coordinates are in the range $1 \dots 1,000,000,000$. Two cows are neighbors if at least one of two criteria is met: (1) If the cows are no further than some integer Manhattan distance C ($1 \leq C \leq 1,000,000,000$) apart. (2) If cow A and B are both neighbors of cow Z , then cow A is a neighbor of cow B . Given the locations of the cows and the distance C , determine the number of neighborhoods and the number of cows in the largest neighborhood.

2. (CF 260C) Andrew plays a game called “Civilization”. Dima helps him.

The game has n ($1 \leq n \leq 3 \cdot 10^5$) cities and m ($0 \leq m < n$) bidirectional roads. The cities are numbered from 1 to n . Between any pair of cities there either is a single (unique) path, or there is no path at all. A path is such a sequence of distinct cities v_1, v_2, \dots, v_k , that there is a road between any contiguous cities v_i and v_{i+1} ($1 \leq i < k$). The length of the described path equals to $k - 1$. We assume that two cities lie in the same region if and only if, there is a path connecting these two cities.

During the game events of two types take place:

- Andrew asks Dima about the length of the longest path in the region where city x lies.
- Andrew asks Dima to merge the region where city x lies with the region where city y lies. If the cities lie in the same region, then no merging is needed. Otherwise, you need to merge the regions as follows: choose a city from the first region, a city from the second region and connect them by a road so as to minimize the length of the longest path in the resulting region. If there are multiple ways to do so, you are allowed to choose any of them.

Dima finds it hard to execute Andrew’s q ($1 \leq q \leq 3 \cdot 10^5$) queries, so he asks you to help him. Help Dima.

9.3.5 Standard Computational Geometry

9.3.6 Less Standard Problems

9.4 Beyond

9.4.1 Data Structure Nonsense