

Advanced Algorithms Simple (for real)



Gabriel Rovesti

Table of Contents

[2 Course Presentation 4](#_Toc161505285)

[3 Graphs 6](#_Toc161505286)

[3.1 Terminology and Concepts 6](#_Toc161505287)

[3.2 Basic Problems, Notations and Properties 8](#_Toc161505288)

[3.2.1 Homeworks 9](#_Toc161505289)

[3.3 Graphs Representation 10](#_Toc161505290)

[3.4 Graphs Algorithms 11](#_Toc161505291)

[3.4.1 Depth-First Search (DFS) 12](#_Toc161505292)

[3.4.2 Breadth-First Search (BFS) 17](#_Toc161505293)

[4 Minimum Spanning Tree 22](#_Toc161505294)

[4.1 Definition 22](#_Toc161505295)

[4.2 Generic Greedy Algorithm 23](#_Toc161505296)

[4.2.1 Theorem and Proof 24](#_Toc161505297)

[4.3 Prim’s Algorithm 26](#_Toc161505298)

[4.3.1 Efficient Prim’s Algorithm – Heap implementation 27](#_Toc161505299)

[4.3.2 Exercises 29](#_Toc161505300)

[4.4 Kruskal’s Algorithm 31](#_Toc161505301)

[4.4.1 Efficient Kruskal 32](#_Toc161505302)

[4.4.2 Union-find Implementation 33](#_Toc161505303)

[4.4.3 Exercise 36](#_Toc161505304)

[5 Shortest Path 37](#_Toc161505305)

[5.1 Single-Source Shortest Path (SSSP) 37](#_Toc161505306)

[5.2 General SSSP Problem 37](#_Toc161505307)

[5.3 All-Pairs Shortest Paths (APSP) 37](#_Toc161505308)

[6 Maximum Flows 38](#_Toc161505309)

[6.1 Maximum Flow Problem 38](#_Toc161505310)

[6.2 Ford-Fulkerson Method 38](#_Toc161505311)

[7 NP-Hardness 39](#_Toc161505312)

[7.1 NP-Hard 39](#_Toc161505313)

[7.2 Cook-Levin Theorem 39](#_Toc161505314)

[7.3 Reductions 39](#_Toc161505315)

[7.4 Maximum Independent Set 39](#_Toc161505316)

[8 Approximation Algorithms 40](#_Toc161505317)

[8.1 Some Algorithms 40](#_Toc161505318)

[8.2 Approximation Algorithm for Vertex Cover 40](#_Toc161505319)

[9 TSP & Metric TSP 41](#_Toc161505320)

[9.1 Travelling Salesperson Problem 41](#_Toc161505321)

[9.2 Metric TSP 41](#_Toc161505322)

[9.3 2-Approximation Algorithm 41](#_Toc161505323)

[9.4 1.5 Approximation Algorithm 41](#_Toc161505324)

[10 Set Cover 42](#_Toc161505325)

[10.1 Greedy Approximation Algorithm 42](#_Toc161505326)

[11 Randomized Algorithms 43](#_Toc161505327)

[11.1 Minimum Cut 43](#_Toc161505328)

[12 Chernoff Bounds 44](#_Toc161505329)

[12.1 Minimum Cut 44](#_Toc161505330)

[12.2 Analysis of Randomized Quicksort 44](#_Toc161505331)

**Disclaimer**

All notes were organized based on the 2022/2023 lessons (considering they are almost identical) and all existing notes files. I tried my best with each, providing precision and refinements over contents, just to give you the clearest idea possible over everything.

The professor is really good and from a pure computer scientist view it’s great in my opinion.

# Course Presentation

(Usual general fluff, then the lectures will be present. This is the only slides-based part, found [here](https://stem.elearning.unipd.it/pluginfile.php/625081/course/section/70344/Advanced%20Algorithms%20-%20Spring%202024.pdf))

Algorithms have a general motivation: create a network of knowledge and allow, with pacing of times, different development and stories creation, while crafting new solutions. We might define them as a sequence of steps to solve the most mundane problems but also really complex ones.

There are different kinds of *applications*:

* network routing
* bioinformatics
* economics (e.g., game theory)
* fluid dynamics
* data mining
* cryptography
* machine learning

The point is this: even when making interviews, algorithms are both the logic and the solution to current problems, thinking *repeatedly and abstractly in a concrete (and fast) way*. Historically, there are still a lot of unsolved or still not found problems. That’s why the course is *mandatory*.

There are also different *goals*, wanting to introduce advanced principles of algorithm design and analysis. In particular, you’ll learn how to:

* Design algorithms for complex domains such as *graphs*
* Recognize “hard” problems and address them using *approximation algorithms*
* Use the power of *randomness* to design fast algorithms
  + and analyze them with appropriate mathematical tools

The *contents* of the course will be the following:

* (Basic) Graph algorithms
  + Graph search and its applications, minimum spanning trees, shortest paths, maximum flows 2 Approximation algorithms
* *Intractable* problems (not solvable in a reasonable amount of time)
  + NP-hardness and reductions between problems
  + Approximation algorithms for intractable problems
    - such as vertex cover, set cover, and the traveling salesperson problem
* Randomized algorithms
  + Main design techniques and analysis tools
    - with applications to problems such as sorting and minimum cuts

Although there are no formal prerequisites, an undergraduate course in algorithms and a good knowledge of (discrete) probability are assumed. Specifically, you should be familiar with:

* *Algorithm design techniques*: divide and conquer, greedy, dynamic programming
* *Data structures*: lists, stacks, queues, binary trees, search trees, heaps
* *Probability*: basic notions, discrete random variables

We want to discuss the *intuition* behind formulating algorithms and distill the core ideas making the algorithms work. Because we are computer scientists, we want to give *rigorousness*: algorithms without proofs are just conjecture and proofs give math logic to those.

We’ll follow, in part, an “active learning” approach:

* Will foster and encourage interaction during class
* Will frequently conclude class with 1-2 exercises
  + whose solution will be shown only at the beginning of next class
* Will frequently post on Moodle further readings
  + news/surveys/research articles/videos related to the topics covered in class
* There is no lab or coding assignments
  + but you are encouraged to code your favorite algorithms up and run them on real data

If you read until here, you sure wanna know: how is the exam?

* Written test, 2 hours. It consists of:
  + 3 questions
    - theory questions on the topics covered in class
    - aimed at verifying the student’s knowledge of the contents of the course
  + 2 problems
    - problems whose solution *requires some creativity*
    - aimed at verifying the student’s ability to use concepts
    - techniques learned during the course to solve new problems

# Graphs

(Suggested readings: The Algorithm, idiom of modern science [[here](https://www.cs.princeton.edu/~chazelle/pubs/algorithm.html)])

A graph is a repartition of the relationships between pairs of objects. In particular, we note:

* as the graph itself
  + = set of vertices (aka nodes)
  + (cartesian product = all) is a collection of edges
    - an edge is a pair of vertices
      * it indicates the connection between two nodes
      * a connection of vertices allows for repetition

In the following drawings, we find:

* directed graphs, which happens if
* undirected graphs, which happens if
* Immagine che contiene calligrafia, schizzo, Line art, disegno

  Descrizione generata automaticamenteImmagine che contiene calligrafia, Carattere, testo, schizzo

  Descrizione generata automaticamentearc = edge inside directed graphs (also called *directed edges*)

In this case, we’ll (mostly) use simple graphs, meaning:

* no parallel edges
* no self-loops

## Terminology and Concepts

We give some *terminology*:

* Given an edge
  + is incident on and (happens if vertex if one of endpoints in that edge)
  + and are adjacent (there is an edge between the two vertices)
* neighbors of a vertex: all vertices s.t.
  + all vertices directly connected to a given vertex by an edge
* degree of a vertex , denoted as or
  + the number of edges incident on

In many ways, graphs are the main modality of data we receive from nature and here we give some *examples*:

* Road networks 🡪 (cities, roads)
* Computer networks 🡪 (computers, computers)
* World Wide Web (WWW) 🡪 (webpages, hyperlinks)
* Social networks (people, friendships relationships)
* Biological networks
  + e.g., molecules (atoms, chemical bonds)
  + e.g., brain (neurons, synapses)
* Finance 🡪 (accounts, transactions)

We give some concepts also:

* path: and
  + finite/infinite sequence of nodes which joins a sequence of vertices via edges
* simple path: (all vertices) are all distinct
  + same definition as above and vertices/nodes are all distinct/so are the edges
  + e.g., has repeated twice so it’s not simple
* cycle: simple path s.t. (starts from a given vertex/ends at same node)
* subgraph:
  + the edges of are incident only on vertices of
  + in words: it is a subset of the larger original graph
* spanning subgraph: a subgraph with
  + a subgraph which “spans” the original graph (so there are all the vertices)
  + following other definitions
    - subgraph obtained by edge deletions only but retaining all vertices
    - so it’s a subgraph of with same vertex set as
* connected graph: if a path from to
* Immagine che contiene diagramma, linea

  Descrizione generata automaticamenteconnected components: a partition of in subgraphs
  + is connected
  + there is no edge between and

Immagine che contiene calligrafia, Carattere, bianco, testo

Descrizione generata automaticamente

Immagine che contiene cerchio, linea

Descrizione generata automaticamente

* tree: connected graph without cycles
  + any two vertices are connected by *exactly* one path

Immagine che contiene cerchio, linea, bianco, appendiabiti

Descrizione generata automaticamenteThere is also the concept of *rooted tree*:

* there is a root
* there is a father for each non-root node and each node is directly linked to the father
* going father to father, we reach

Continuing with definitions:

* forest: set of trees (disjoint)
  + also = undirected graph in which any two vertices are connected by *at most* one path
* Immagine che contiene schizzo, calligrafia, disegno, Carattere

  Descrizione generata automaticamentespanning tree: a spanning subgraph connected and without cycles
* Immagine che contiene schizzo, cerchio, design

  Descrizione generata automaticamentespanning forest: a spanning subgraph without cycles

## Basic Problems, Notations and Properties

There are different *basic problems*:

* Traversal (systematic exploring of graph e.g., crawling)
* Connectivity (tell if the graph is connected or not)
* Computing connected components (e.g., wireless networks)
* Spanning trees (e.g., efficient broadcasting in wireless networks)
* Minimum-weight spanning trees (e.g., navigator)
* Shortest paths (e.g., social media friend analysis)

Also consider some notations and properties:

* (number of nodes)
* (number of edges)
* the size of this graph is
  + is not enough (normally online you would find the size it’s = count of edges)
  + consider a scenario of a graph with vertices and no edges
  + the size of graph would be but we don’t consider vertices
    - we are accounting for both the “space” and the “connections” occupied

### Homeworks

Exercise (Properties of graphs)

*Let be a simple, connected graph with vertices and edges. Then:*

1. *is a tree*
2. *is connected*
3. *is acyclic (i.e., is a forest)*

*Prove the previous properties*.

My solution

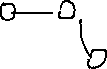
Consider an example for each one:

1. . The degree is the number of vertices incident in . The example clarifies it.



This happens because inside the summation, every edge is counted twice.

1. Consider vertices. This happens because we are choosing vertices out of to form an edge. In a simple graph, order does not matter, so we can select any two vertices, then arranging with all the possible arrangements and avoid counting each pair twice. Indeed, in a simple graph there are possible pairs of vertices.
2. If is a tree (=connected graph without cycles), consider and, you would have:



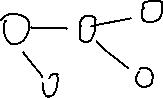
Fix a root. Then, represents father-child relationships, which are .

1. If is connected we would have



This happens because is a tree that may have cycles, thus it can only have more edges.

1. Consider an acyclic graph, then



This happens because is a tree that may not be connected, thus it can only have less edges.

Official solution

1. In the summation, every edge is counted exactly twice
2. In a simple graph, there are possible pairs of vertices
3. Fix a root on a vertex (so, consider as rooted tree, thanks to the equivalence between rooted tree and “free” tree). Then represent father-child relationships, which are (which means each non-root node has a unique father)
4. is a tree that may have cycles it can only have more edges than a tree
   1. Consider connectivity removes edges and keeps the graph connected without cycles, thanks to edges
5. is a tree that may not be connected it can only have less edges than a tree
   1. If it is a tree without cycles, it is a forest, and its maximum edges are

## Graphs Representation

How to encode a graph for use in an algorithm?

Consider a list of vertices and a list of edges . (they contain all information about and and the links between each other). Let’s consider vertices are called . This is useful but does not allow for fast algorithms overall.

To allow for *direct access to edges*, one of the following data structures are used, in addition to pointers to , .

* an adjacency list
  + an array of lists, one vertex (consider the example below)
  + each containing all the vertices adjacent to (represented by table below)

|  |  |
| --- | --- |
| 1 | 2,5 |
| 2 | 1,3,4,5 |
| 3 | 2,4 |
| 4 | 2,5,3 |
| 5 | 4,1,2 |

Immagine che contiene cerchio, linea, diagramma, schizzo

Descrizione generata automaticamente

What if directed? Only vertices pointed for that vertex.

* Pro: space usage i.e. linear
* Con: no quick way to determine if a given edge is in the graph
* an adjacency matrix
  + a matrix s.t. if , otherwise



Immagine che contiene schermata, numero, Carattere, diagramma

Descrizione generata automaticamente



* If graph is directed 🡪 the matrix is *asymmetric*
* If graph is undirected 🡪 the matrix is *symmetric*
  + edges are bidirectional 🡪 only half of matrix needs to be stored
  + operations here are more efficient in general

In case of a *weighted graph*, each cell of the matrix has either the value of the edge weight (as number) or to represent null costs. This kind of graph represents costs, capacities, etc.

* Pro: Quick to determine if a given edge is present
* Con: Space required is 🡪 can be superlinear in the input size
  + if number of vertices increases, the space required by matrix grows quadratically

It may also depend on the number of edges:

* *dense* graph = number of edges close to maximal number
  + many cells inside adjacency matrix will be populated by non-zero values
  + adjacency matrix is mostly used here
    - allows to quickly test the presence of an edge and check its info
* *sparse* graph = number of edges with only a few edges
  + conversely, majority of values will be zero

## Graphs Algorithms

We are focusing over *graph search and its applications*, in particular traversal/exploration. They provide a systematic way to explore a graph starting from a vertex ( source vertex) visiting all the vertices (starting from a graph and a source vertex). The most famous algorithms are:

* *Depth-First Search (DFS)* 🡪 aggressive, goes in depth, then comes back and so forth
* *Breadth-First Search (BFS)* 🡪 non-aggressive, proceeding by levels inside graph

Immagine che contiene schizzo, disegno, cerchio, clipart

Descrizione generata automaticamenteImmagine che contiene schizzo, diagramma, cerchio, disegno

Descrizione generata automaticamenteIn particular, consider the following graphs; in each, the types of visits are defined already in color.



Observe that:

* DFS and BFS serve as design patterns, acting as building blocks
  + where the visit operation can be instantiated to solve specific problems
  + such as connectivity and spanning tree identification
* Traversing and lists also achieves complete graph exploration
  + however, the lack of systematic exploration makes it less useful for problem-solving
* The idea behind both is to prioritize visiting neighbors with lower IDs from the starting vertex

### Depth-First Search (DFS)

(Further readings for this one: [paper](https://stem.elearning.unipd.it/pluginfile.php/749331/mod_folder/intro/Depth-First%20Search%20and%20Linear%20Graph%20Algorithms%20-%20Tarjan%20%281972%29.pdf) and [survey](https://arxiv.org/abs/2201.07197))

This is a recursive algorithm which:

* starting from a source “visits” all vertices of the connected component containing
* uses adjacency list as graph representation
* every vertex has a field which can be either
  + if visited
  + otherwise
* every edge has a label which can be either
  + initially
  + or
    - first label indicates an edge which allows discovery of vertices
    - second label indicates non-tree edges
      * that go from a node in the DFS tree
      * to some ancestor of in the DFS tree
      * this kind of edges is useful in order to find cycles

Consider the following procedure:

(*first invoke: )*



Because I like people understanding stuff, let’s comment human-like this code, considering we:

* take each vertex and we see if it was visited or not
  + this is done *on the connected component* touching all vertices and edges
  + we use adjacency lists to induce an order of visit in neighbors
* check if the current vertex (with ID field) has been visited or not
* loop on all edges incident to current vertex
* check if label of current edge was not labeled = it was not explored
* consider the opposite vertex = other endpoint of the edge
* if that opposite vertex has not been visited yet
  + edge leads to an unexplored vertex, indicating a discovery edge
    - a vertex is discoverable if there exists a path between and not visited
  + this will be labeled, indicating it’s the first time the edge is being traversed
  + then we recursively call the algorithm to explore the connected component
* else (aka it was already visited)
  + the edge is leading to an already explored vertex
  + the edge is labeled indicating a connection back to the ancestor

Immagine che contiene diagramma, cerchio, linea, disegno

Descrizione generata automaticamenteThe following is an example of the algorithm being applied:

#### Correctness

At the end of the algorithm:

1. all the vertices of have been visited and all the edges in are labelled either
2. the set of is a spanning tree of called “DFS tree”

*Proof*:

1. (short: by construction)

By contradiction, not visited. Since is connected, there is a path from to

. Let be the first unvisited vertex in the path (

We run into the contradiction: must have been executed and therefore is called (meaning was found not visited). This happens in contradiction to the hypothesis (it’s not possible to find a vertex unvisited and marked as such).

A vertex is visited only when is invoked DFS is called all incident edges on are labelled, by construction.

1. DFS is called , once, and , a vertex s.t. and is labelled and is invoked from . We say that gets “discovered” by and let’s call “father” of
   1. father (there exists a father and it is unique)
   2. going back father to father eventually is reached

Then, the set of is a *rooted tree* that *touches* all the vertices of and it’s a spanning tree of (unique path from every vertex to the source one and each is discovered by exactly one parent vertex).

#### Complexity

Given:

* : number of vertices of (one invocation )
* : number of edges of (costs related to node, excluding recursive invocations inside)

The complexity overall is:

(sum of degrees of each node proportional to number of edges)

Note that is connected, so:

* (connected, so for vertices we would have at least edges)
* (n. of edges at least proportional to n. of vertices)

#### Extension

The possible extension is to visit all the graph (aka: all components even if not connected):

Overall, the complexity if because it scans over all the vertices and nodes.

#### Homeworks

1. *Given a graph and two vertices determine, if it exists, a path from to*
2. *Given a graph return a cycle (if any)*

My solution

1)

2)

Official solution

* 1st exercise ( path)
  + add a field .
  + Modify s.t. when a is labeled
    - then
  + Run . Check if has been visited
    - NO: then return “No path”
    - YES: starting from , follow the “parent” label, so as to build a path from to
  + Complexity: where is the number of edges of connected component
* 2nd exercise (cycle) 🡪 we go back thanks to back edges because they “close” the cycles
  + add a field and add a field
  + is a then
  + is a then
    - then is an ancestor of in the DFS tree
  + Run DFS on each connected component
  + Check all the edges
    - as soon as an edge is found as
    - and
    - then return a cycle adding to all the edges found in the path from to
    - if no is found, then return “No Cycles” (it would be a tree)

Complexity for both algorithms: 🡪 invoked DFS once for each connected component

#### More applications

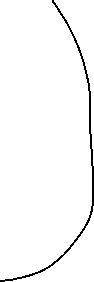
More or less what we did until now with DFS was returning a spanning tree. Other problems which can be solved are the following:

* *graph connectivity*: return whether the graph is connected or not
* *connected components*: return a labeling of all the vertices of s.t. 2 vertices have the same label if and only if they are in the same connected component

We will modify the algorithm in such a way that (let’s generalize it, so) (integer, label of the k-th component)

*Graph connectivity*

*Connected components*



Complexity of the whole thing:

#### Summary

Given a graph , the following problems can be solved in using DFS:

* test if is connected
* find the connected components of
* find a spanning tree of (if is connected – otherwise it’s called spanning forest)
* find a path between two vertices (if any)
* find a cycle (if any)

Possible questions for exam: Show how to find a spanning tree in linear time/Find if graph is connected in linear time, something like that.

### Breadth-First Search (BFS)

This is an iterative algorithm that starting from a source vertex “visits” all the vertices in the same connected component of , and partitioning the vertices in levels depending on their distance from (with distance we mean the shortest path). We’ll use adjacency list to represent :

* if not visited, if visited
* if has no label,
  + connect vertices at different levels (different labels)

An explanation step-by-step of the algorithm:

* visit the source vertex
* iterate over all levels
* explore neighbors starting from level
* create a set of vertices for the next level
* for all incident edges
  + if vertex has not been visited
  + we get the opposite vertex
  + if vertex has not been visited yet
    - mark edge as
    - visit vertex
    - mark it as visited
    - add the vertex to the set of vertices for the next level
  + else (if vertex has already been visited)
    - node represents crossing between different levels
    - it will be marked as
    - it connects two nodes that don’t share any ancestor-descendant relation
  + increment the level counter
  + algorithm terminates where there are no more vertices to visit

Immagine che contiene disegno, schizzo, Arte bambini, diagramma

Descrizione generata automaticamenteHere we do the following example (at first invoke: :

Immagine che contiene cerchio, diagramma, schizzo, linea

Descrizione generata automaticamente

The algorithm, when executed, it will behave like the following:

* + takes source vertex and marks it as visited
  + goes to next level and for each vertex takes its connected component (incident edge)
  + so, from goes to because it’s connected
    - given it was not explored yet, also jumps to
  + same thing for and
  + they were the connected components of and , jumping to the opposite edge
  + from we see a connected component, found within next level of adjacency list
  + given they were already explored, between two are discovered

#### Correctness

At the end of we have:

1. all vertices in are visited and all edges are labelled
2. the set of are a spanning tree (tree touching all the vertices) of
   1. analogously to DFS, we call it BFS tree (this is rooted in )
3. the path in from to has edges and every other path from to has exactly edges (e.g., edges)

Proof of (1) and (2): same as for the DFS

Proof of (3):

* let where is “discovered” from is a is a path of
* By contradiction, assume a path : with (shorter)
* This implies that … or or … (might be on some levels before)
* This means that, since but this is a contradiction

In words:

* what you have is that the first node is in level zero, so on and so forth until the tth node is in some list between and
* if this were true, then we have that , because we have that
* this is absurd since we assumed that

#### Complexity and applications

there is one iteration of the first for loop and iterations of the second for loop.

Complexity: (which becomes if is connected; in general, each for execution and access to lists).

Some *applications*:

1. same as for DFS in time
2. given a graph and return the *shortest* path (in terms of least amount of edges between two nodes) from to (if any)
   1. modify s.t. when is labeled then
   2. run and return the set of child-parent edges
   3. Complexity:

*Solutions*:

1)

Basically, if we explore all vertices and incident edges we would get a complexity of . This is infact what we are doing here.

2)

Explanation of this algorithm:

* starting from source vertex
* visit it and mark it as such
* create a list of vertices containing the source one
* for all vertices of list and for all incident edges to that vertex
* if vertex was not labelled yet
  + it is marked as
  + parent of becomes and it is now possible to go up
  + it is marked as visited
  + it’s inserted into the list
* otherwise, we mark it as cross edge

Complexity: where is the number of edges of connected component .

# Minimum Spanning Tree

The goal is to *interconnect* a set of objects in the cheapest possible way (e.g., connecting PCs inside departments using the least amount of cable as possible). It’s a fundamental problem, studied since the 1920s.

## Definition

More specifically, its *definition* is the following:

* Input: a graph undirected, connected and *weighted*
  + A weight
  + defines cost of edge
* Output: a spanning tree of s.t. is minimized
  + Goal is minimizing the sum for all weights of every edge of the tree

Immagine che contiene disegno, schizzo, clipart

Descrizione generata automaticamenteConsider the following example; here, the MST is made of the blue part (minimum part to cover all vertices – for the sake of simplicity, we consider it starting from ):

We give the following observations:

* minimum-~~weight~~ spanning tree
  + not to be confused with e.g. minimum number of sides
  + because *all* spanning trees have the same number of sides
    - and we want the one that weighs the least (it may not be unique)
* connected assumption is without loss of generality (wlog)
  + if graph is not connected, we talk about Minimum Spanning Forest ()
    - = MST for each connected component

There are different *applications* we can define:

* Networks (computers, sensors, electrical)
  + E.g., broadcast determining a backbone
    - a subgraph connecting all network nodes and with minimum cost
* Machine learning (building block for clustering algorithms)
* Computer vision (object detection)
* Data mining
* Subroutine in other (approximation) algorithms
  + To solve other problems

We ask ourselves some *questions*:

* How difficult is it?
* How many spanning trees can a graph have?

Immagine che contiene schizzo, modello, design, arte

Descrizione generata automaticamente con attendibilità mediaThe simplest MST algorithm is to compute all the spanning trees and select the one with minimum weight. We would need a *complete graph* to do that: it has all possible edges.

A complete graph has different spanning trees (worst case would be exponential, quantity larger than the universe, even on the small graphs this would not work). The right figure shows a complete graph.

However, surprisingly, MST can be solved in near-linear time (specifically in ). It can be done using *greedy algorithms* simpler to understand and implement in practice (e.g., Prim, Kruskal). They both apply (in different ways) a *generic greedy algorithm*.

## Generic Greedy Algorithm

The idea of a generic-MST algorithm is to maintain the following invariant:

* at each iteration, is a subset of edges of some MST
* every time an edge is chosen, this is considered to be a right edge
  + because the choices are made in a greedy manner
* at each iteration, the algorithm adds an edge that does not violate the invariant
  + considered “safe” edge for (safe to add it/don’t do a mistake if you do)

We give the algorithm then:

This is simple but does not say anything on how to find a “safe” edge. So, the question is exactly:

* How to find a safe edge?

Luckily, MSTs enjoy the structural property given by the Theorem next. First, some *definitions*:

* A cutof graph is a partition of
  + in words, a partition of vertices into two disjoint subsets
  + it can be done on one or more edges
* An edge crosses a cut if and (or viceversa)
  + so, if its endpoints lie in different subsets of the partition defined by the cut
* A cut respects a set of edges if no edge of *crosses* the cut
* Given a cut, an edge that crosses the cut and is of minimum weight is called light edge(for that cut) 🡪 they are useful, because when included in MSTs, they have minimum weight

We give the following example; a simple cut on three edges and the light edge is to be considered as such because it’s respectful - it doesn’t cross the cut and it’s the minimum weight:

Immagine che contiene calligrafia, disegno, Arte bambini, design

Descrizione generata automaticamente

### Theorem and Proof

*Theorem*:

Let be an undirected, connected and weighted graph. Let be a subset of included in some MST of , let a cut that respects and let be a light edge for . Then is safe for .

Immagine che contiene diagramma, schizzo, linea, design

Descrizione generata automaticamenteConsider the following example of ; it basically just throws cuts at random and selects the edge with minimum cost which “respects” the others (aka it was not taken before), connecting possibly all vertices at least once, because it has to “span” them:

As said, this one considers cuts at random; other algorithms have rules to choose cuts (e.g., Kruskal).

*Proof of theorem*:

It uses the technique of “cut and paste”, standard technique used in the context of greedy algorithms.

* Cut-and-Paste is a way used in proofing graph theory concepts
  + Idea is this:
    - Assume you have solution for Problem , you want to say some edge/node, should be available in solution
    - You will assume you have solution without specified edge/node
    - You try to reconstruct a solution by cutting an edge/node
      * and pasting specified edge/node
      * and say new solution benefit is at least as same as previous solution
* So: fake to take an optimal solution and transform it in the solution returned by the algorithm
* This shows the cost of the two solutions is the same
  + and also the solution returned by the algorithm it’s optimal
* One of most important examples is proving MST attributes

Immagine che contiene disegno, schizzo, Line art, clipart

Descrizione generata automaticamenteLet be an MST that includes (basically, we take an optimal solution considering safe edges). Assume that (otherwise, we’d be done). We’ll build an new MST that includes . Consider the following example (it simply considers an MST with and not ):

By hypothesis, crosses another edge of that crosses that cut (it would exist because it’s a spanning tree, if there wasn’t it wouldn’t be in the first place).

* By hypothesis, respects removing from and adding we obtain a new spanning tree that includes
  + because it doesn’t partition it and simply adds regularly
* Now we need to show that not only is a ST (spanning tree), but also a MST



* and both cross but by hypothesis is the light edge between the two
* But is an MST

In words: we've shown that adding an edge between vertices to a tree, making it a MST again, maintains its optimality. By proving that the added edge was already a greedy choice, we've shown that its inclusion in the graph does not increase the weight of the tree and it maintains its properties safely.

We’ll now see two MST algorithms that organize these “respectful” choices.

## Prim’s Algorithm

This algorithm was crafted in 1957. How does Prim’s algorithm apply :

* is a single tree
* safe edge: a light edge that connects the tree with a vertex that does not belong to the tree

Here goes the pseudocode:

Immagine che contiene disegno, calligrafia, schizzo, diagramma

Descrizione generata automaticamenteWe explain with a figure how this algorithm works:

In summary, Prim's Algorithm operates by iteratively selecting safe edges, so light edges connecting the current tree with vertices not yet included in the tree.

* Basically, it includes two sets of vertices
  + the ones already included in the MST
  + the other containing the ones not yet included
* At every step, it considers all the edges that connect the two sets and picks the minimum weight edge from these ones
* After picking that one, it moves the other endpoint of edge containing the MST

This algorithm “*grows*” a spanning tree from a source vertex (doesn’t matter who is) by adding an edge that a time. This is implemented resulting in an efficient and optimal solution.

Immagine che contiene diagramma, linea, calligrafia, disegno

Descrizione generata automaticamenteIn the following example, a random vertex is chosen and then the algorithm is applied, following the principles discussed before in various points:

* *Correctness*: it follows from the [theorem](#_Theorem_and_Proof)
* *Complexity*:
  + iterations doing at most vertices
  + assuming is represented with an adjacency list
  + polynomial time 🡪 efficient algorithm (for small graphs)
    - this time is not efficient in practice (big/large/very large graphs)
      * think of Facebook graph:
      * so 🡪 not so efficient in very large graphs

### Efficient Prim’s Algorithm – Heap implementation

Key observation: in the basic implementation, the computation of the light edge (minimum) is done *repeatedly* should speed it up, so the algorithm and the computation will be faster.

Golden rule in algorithms/coding: when an algorithm repeats frequently the same question, look for the “right” data structure to speed that operation up. The right kind of data structure to do this is a *priority queue*, implemented with a heap.

* Recap about this data structure
  + 🡪 add an object to the heap (possibly fast)
  + 🡪 remove an object with the smallest key (highest priority)
  + 🡪 given a pointer to an object, remove it
* In a heap with objects, the complexity of these operations is

We can redefine the Prim’s algorithm exploiting this efficient data structure, basically with the same principle; consider a min heap starting from whatever vertex, which is the root. From there, always extract the minimum value (means checking if it is min heap), then update the path.

Below figure shows that:

Immagine che contiene disegno, diagramma, schizzo, calligrafia

Descrizione generata automaticamente

It’s simple to store *vertices* in the heap (instead of edges) – whatever operation, the heap property will be established again, restoring keys and values. We can redefine Prim’s implementation with a heap:

We consider, during algorithm execution implicitly, we have

*Complexity*:

* 🡪
* 🡪 iterations
* 🡪

Total cost of only operations:

* *for loop*: executed times in total (every vertex is explored)
  + - this here is a simple check
    - two operations

Total cost of *for loop*: (iterating for all adjacent nodes, quantity equal to node degree)

This way, the total complexity of the algorithm is (since is connected, we recall) 🡪 near-linear time.

### Exercises

Exercise (uniqueness of MSTs):

*Show that if the weights of the edges are all distinct then there exists exactly one MST.*

(*Hint: cut and paste argument – similar to the theorem correctness*)

My try on the solution

Immagine che contiene linea, schizzo, diagramma, disegno

Descrizione generata automaticamenteWe consider for the sake of example two MSTs, which we assume to be the contrary one to the other:

The copy and paste argument would suggest removing one edge from the first MST and add it to the second one, for example. Doing so would require adding a cycle to the original one, given the weights are all distinct and doing this operation would require making a loop between existent edges.

Immagine che contiene design, bicicletta

Descrizione generata automaticamente con attendibilità bassaOfficial solution (with details of lesson but also other including Wikipedia and other sources)

Assume there are two MST different from each other, so the contrary and so an edge in one but not in the other; since weights are distinct, with min weight, call it , without loss of generality (not introducing any artificial assumption), and the argument is a cut-and-paste one (this choice will be unique, considering edge weights are all distinct from each other):

* add to this creates a cycle ; is (M)ST no cycles has an edge
  + because was chosen as the unique lowest-weight edge (only edge with minimum weight not in the other) among those belonging to exactly one of and
  + therefore the weight of must be greater than the weight of
* remove from get a new spanning tree with weight (so, smaller weight): contradiction, because is an MST!

Two conclusions can be done:

* more generally, if the edge weights are not all distinct then only the (multi-)set of weights in minimum spanning trees is certain to be unique; it is the same for all minimum spanning trees
* when the edge weights are not all distinct, it's possible for multiple different MSTs to exist
  + however, while the actual arrangement of edges in these MSTs may vary, the set of weights of the edges across all MSTs will remain the same

Exercises

1. *Is the converse true? (e.g., are weights necessarily unique for every possible graph and this has to hold for every graph)*

Revised solution (considering the very few sources online)

A connected graph with repeated edge weights and this can still have a unique minimum spanning tree. Considered the trivial example of being a tree; in this case, there are no cycles, so any spanning tree will be minimal, hence unique, regardless of repeated edge weights.

In conclusion, we might say:

* Distinct weights guarantee a unique MST
* Repeated weights can have multiple MSTs
  + but the set of weights used will always be the same across all of them

1. *Show that the second best MST, that is, the spanning tree of second-smallest total weight, is not necessarily unique (here we look for only one graph)*

Considering a minimum spanning tree is connected to at most edges, we would have at least an edge not in the minimum spanning tree. They will have necessarily, wherever we have exactly one edge, a situation where the weights are equal.

Proving there exist one, needs the assumption of the weights to be distinct. Adding an edge would form a cycle and here the weight of the new edge would be, reestablishing the MST property, again another MST with an edge of minimum weight. We would from a tree with a cycle having weight lesser than the original tree and it’s not the original tree.

Solution from the book itself (coming from [here](https://viterbi-web.usc.edu/~shanghua/teaching/Spring2010/public_html/files/HW2_Solutions_A.pdf))

Since any spanning tree has exactly edges, any second-best minimum spanning tree must have at least one edge that is not in the (best) minimum spanning tree. If a second-best minimum spanning tree has exactly one edge, say that is not in the minimum spanning tree then it has the same set of edges as the minimum spanning tree, except that replaces some edge, say , of the minimum spanning tree. In this case, , as we wished to show.

Thus, all we need to show is that by replacing two or more edges of the minimum spanning tree, we cannot obtain a second-best minimum spanning tree. Let be the minimum spanning tree of , and suppose that there exists a second-best minimum spanning tree that differs from by two or more edges. There are at least two edges in and let be the edge of with minimum weight. If we were to add to , we would get a cycle . This cycle contains some edge in (since otherwise would contain a cycle).

We claim that We prove this claim by contradiction, so let us assume that . (Recall the assumption that edge weights are distinct, so that we do not have to concern ourselves with ). If we add to we get a cycle , which contains some edge in . (since otherwise, would contain a cycle). Therefore, the set of edges forms a spanning tree, and we must also have , since otherwise would be a spanning tree with weight less than. Thus, which contradicts our choice of as the edge in of minimum weight.

Since the edges and would be on a common cycle if we were to add to , the set of edges is a spanning tree, and its weight is less than . Moreover, it differs from (because it differs from by only one edge). Thus, we have formed a spanning tree whose weight is less than but is not . Hence, was not a second-best minimum spanning tree.

## Kruskal’s Algorithm

Another very simple, very famous and fast algorithm for MST is Kruskal, which dates back to 1956.

* It’s as fast as Prim, both in theory and in practice (if properly implemented)
  + very quick and clean in its implementation
* It gives us the opportunity to study a new data structure

It implements the algorithm, but:

* is a forest (set of trees)
* safe edge is a light edge connecting distinct components

Here goes the pseudocode:

Immagine che contiene diagramma, linea, design

Descrizione generata automaticamenteConsider the following example – vertices will be sorted by weight and the result is not a tree but a set of two trees, adding only ones that do not form a cycle:

Simple optimization of the algorithm: stop the *for loop* when has edges. Note also that a source vertex is not needed here.

*Correctness*: follows from correctness of ([here](#_Theorem_and_Proof) – basically, it’s always the same algorithm, only seen in a different way, considering it’s always “respecting” cuts in a safe way, as shown in next figure).

Immagine che contiene schizzo, disegno, diagramma, linea

Descrizione generata automaticamente

*Complexity*:

* sorting:
* for loop: check whether closes a cycle is equivalent to check whether contains an path 🡪 DFS on 🡪 complexity:

Total: 🡪 [because implemented with adjacency list]

Question: *Can we implement it faster?*

### Efficient Kruskal

It can be implemented as fast as Prim’s, considering only frequent operations done inside the previous algorithms.

* Kruskal frequent operation = cycle check (equivalently, path check)
  + happens when an edge is added to

We create a new data structure supporting this operation fast and to do that, we use a structure called Union-Find (also called *disjoint set*), created in 1964. This is a structure to merge *disjoint sets* (also non-overlapping in their elements) of objects and supports at least three operations:

* *Init*: given an array of objects
  + it creates a Union-Find data structure with each object in its own set
* *Find*: given an object , return the name of the set that contains
* *Union*: given two objects merge the sets that contain and into a single set
  + done whenever the sets are distinct
  + if are already in the same set, this operation does nothing

These operations can be implemented with the following complexities:

* *init*:
* *find*:
* *union*:

(here, we assume represents the number of objects in the data structure)

We want to implement Kruskal fast using the Union-Find structure:

* Idea: Union-Find (U-F) keeps track of the connected components of the current solution
  + whenever I need to check if the merge unites with existing edges
  + creates a cycle are already in the same connected components

*Complexity:*

* Init:
* Sorting:
* Find:
* Union: 🡪 only when I go inside an “if” and when the edge is added
* updating:

In total:

### Union-find Implementation

(Further readings for this part: [here](https://stem.elearning.unipd.it/pluginfile.php/749335/mod_folder/intro/eisner.mst-tutorial.pdf), [here](https://stem.elearning.unipd.it/pluginfile.php/749335/mod_folder/intro/The%20saga%20of%20minimum%20spanning%20trees.pdf) and [here](https://www.cs.cmu.edu/afs/cs.cmu.edu/academic/class/15859-f09/www/handouts/cmuonly/graham-hell.pdf))

We’ll use an array, which can be visualized as a set of directed trees. Each element of the array has a field that contains the index of the array of same object .

Immagine che contiene testo, schermata, Carattere, linea

Descrizione generata automaticamenteConsider the following example:

Other things:

* vertices: (indexes of) objects
  + called by index inside of the array
* arc

Immagine che contiene calligrafia, disegno, schizzo, Line art

Descrizione generata automaticamenteA set of object corresponds with a set of trees directed to the “parent graph”:

As convention: name of the set = root of that tree. We are giving the names of the roots, hence why we call them and , visualized with direct trees. Consider the following operations:

* Immagine che contiene schizzo

  Descrizione generata automaticamenteInit

Allows to initialize nodes as the parents of themselves (hence, the self-loops you see above)

* Find

It follows parent by parent until the root is reached

Consider the following procedure – finds the name of a set containing the node to find and returns it. Basically, goes up father by father reaching a root identifiable as a self-loop

1. starting from ’s position in the array
   1. traverse parent edges until reaching a position
   2. s.t.
2. return

*Definition*: The depth of an object is the number of edges traversed by

*Examples*:

The *complexity* *of*  is the largest depth of , depending on the Union implementation.

* In general, the worst case complexity is , depending on how the trees are merged
* In any case, the maximum height of a tree is

Immagine che contiene calligrafia, testo, Carattere, tipografia

Descrizione generata automaticamenteAs shown in figure, this strictly depends on the specific case; some are better, some are worse:

* Union

given two objects

* the two trees of the parent graph containing and must be merged in a single tree
* the simplest way is to point one of the 2 roots to one *another node* of the other tree

We need to decide:

1. which of the 2 roots remains a root
2. Immagine che contiene schizzo, disegno, Line art, Arte bambini

   Descrizione generata automaticamenteto which node should a root point

In the right figure (for Point 1) the root (4) is pointed, because we want to keep the depth as low as possible. That’s the simplest thing that we can do.

Every level below the root would make the depth bigger by 1. That’s why the others in the figure are considered bad.

(For Point 2) We would need to have (6) pointing (4), because otherwise depth would increase – minimum possible increasing depth.

We would need to have the smallest amount of nodes pointing to the biggest one – in other word, to have the leas increasing depth, there are two alternatives, concerning Point 1:

* One possible idea for is *to minimize the number of objects* whose depth will increase
  + this way, objects depth increases as much as the depth of the tree in which it’s been chosen to set the tree
  + this is called *union-by-size*. This is the chosen option indeed.
* Alternatively: the root of the less-tall tree points to the one of the tallest tree
  + this is called *union-by-rank*

For the procedure :

1. invoke and to obtain the names and of the sets that contain and
   1. basically, just to understand in which tree they are
   2. if then (aka – there is nothing to do)

### Exercise

Exercise

*Argue that the complexity of (and of ) is*

My solution

Initially, clearly, root has depth 0. Then, we are uniting trees, so starting from the Union operation itself, we link the root of the other tree, this way considering at least two nodes each time. Given the depth increases linearly, the size of the tree gest doubled each time. So, considering the size of the tree must be respected, we should have a complexity of , given the depth of the tree gets doubled until maximum height of said tree (remember this is .

Official solution

Initially, . can only increase because of a Union in which the root of the tree of points to another root. This happens only when the tree of gets merged to a tree of size not smaller when the depth of increases, the size of the tree of at least doubles. This can happen specifically times, therefore the depth of cannot increase more than times.

So, we have two different algorithms with complexity . To reach complexity is still an open problem.

# Shortest Path

## Single-Source Shortest Path (SSSP)

aaa

## General SSSP Problem

aaa

## All-Pairs Shortest Paths (APSP)

aaa

# Maximum Flows

## Maximum Flow Problem

aaa

## Ford-Fulkerson Method

aaa

# NP-Hardness

## NP-Hard

aaa

## Cook-Levin Theorem

aaa

## Reductions

aaa

## Maximum Independent Set

aaa

# Approximation Algorithms

## Some Algorithms

aaa

## Approximation Algorithm for Vertex Cover

aaa

# TSP & Metric TSP

## Travelling Salesperson Problem

aaa

## Metric TSP

aaa

## 2-Approximation Algorithm

aaa

## 1.5 Approximation Algorithm

aaa

# Set Cover

## Greedy Approximation Algorithm

aaa

# Randomized Algorithms

## Minimum Cut

aaa

# Chernoff Bounds

## Minimum Cut

aaa

## Analysis of Randomized Quicksort

aaa