CS300 Algorithms

Graph Algorithms

What is a graph?

Definition A *graph* G consists of a set $V(G)$ called *vertices* together with a collection $E(G)$ of pairs of vertices. Each pair $\{x, y\} \in E(G)$ is called an *edge* of G.

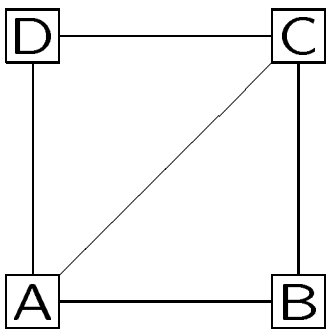
Example If

$$V(G) = \{A, B, C, D\}$$

and

$$E(G) = \{\{A, B\}, \{C, D\}, \{A, D\}, \{B, C\}, \{A, C\}\}$$

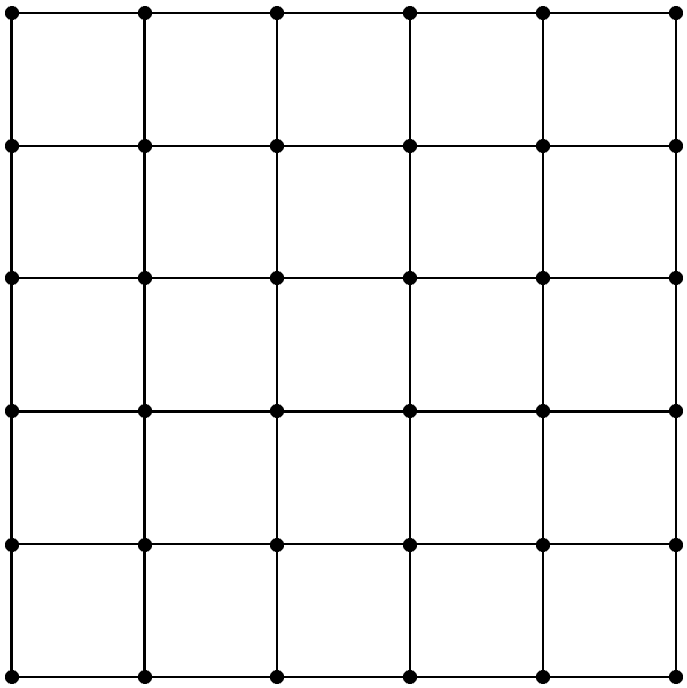
then G is a graph with 4 vertices and 5 edges.



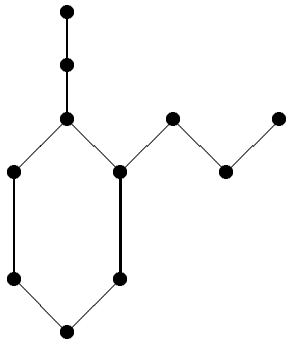
What are graphs used for?

Graphs are used to model any kind of *binary relationship* in many different domains. Here are some examples:

In computing: The vertices of the graph are processors in a parallel computer. The edges connect processors that are directly joined by a communication link.



In chemistry The vertices are carbon atoms in a molecule, and there is an edge between two vertices if there is a bond between the corresponding atoms.

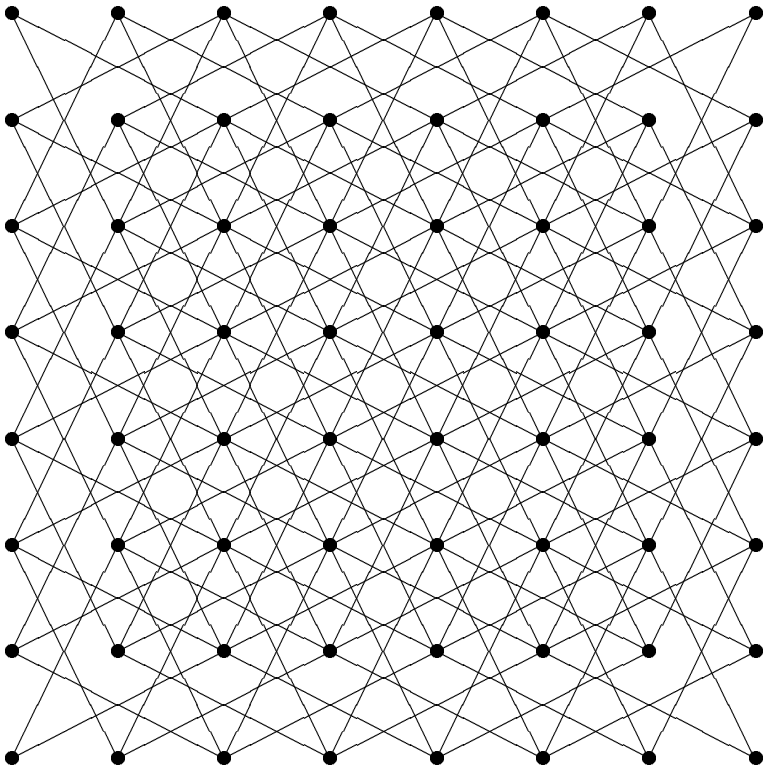


In mathematics The vertices of the graph are the points of the Euclidean plane. Two vertices are joined by an edge if the distance between them is exactly one.

This graph is hard to draw because it is an *infinite graph*.

More examples

In games The vertices are the 64 squares on a chessboard, and two vertices are joined by an edge if a knight can move from one square to the other.



The problem of finding a *knight's tour* is equivalent to finding a *Hamilton cycle* in this graph. (Hamilton cycles will be defined later.)

Six degrees of separation

The Acquaintanceship graph The vertices of the graph are all the people in the world. Two vertices are joined by an edge if the corresponding people know each other by name.

It is often stated that there are at most six links between any two people in the world—in graph theory this is equivalent to saying that the *diameter* of the acquaintanceship graph is six.

Erdös number Consider a graph whose vertices are all the authors who have published a paper in a refereed journal. There is an edge between two vertices if the corresponding authors have published a joint paper in a refereed journal.

The *Erdös number* of a mathematician M is the *distance* of M from the vertex "Erdös".

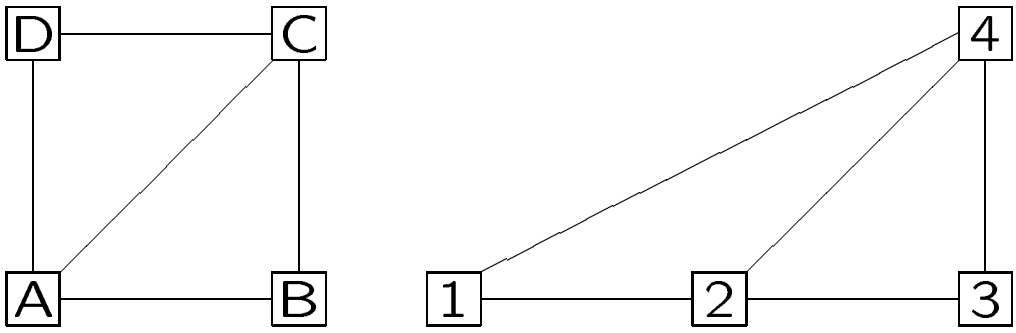
See

`http://www.acs.oakland.edu/~grossman/Erdoshp.html`

(linked from ALG page) for details about this graph.

Isomorphisms

Consider the following two graphs:



Apart from the "names" of the vertices and the geometric positions it is clear that these two graphs are basically the same — in this situation we say that they are *isomorphic*.

Definition Two graphs G_1 and G_2 are *isomorphic* if there is a one-one mapping $\phi : V(G_1) \to V(G_2)$ such that $\{\phi(x), \phi(y)\} \in E(G_2)$ if and only if $\{x, y\} \in E(G_1)$.

In this case the isomorphism is given by the mapping

$$\phi(A) = 2 \quad \phi(B) = 3 \quad \phi(C) = 4 \quad \phi(D) = 1$$

The `Graph` interface

There are many ways to implement graphs, and graph-theoretic algorithms in Java. Using an interface to represent a "bare-bones" graph seems to be the most useful.

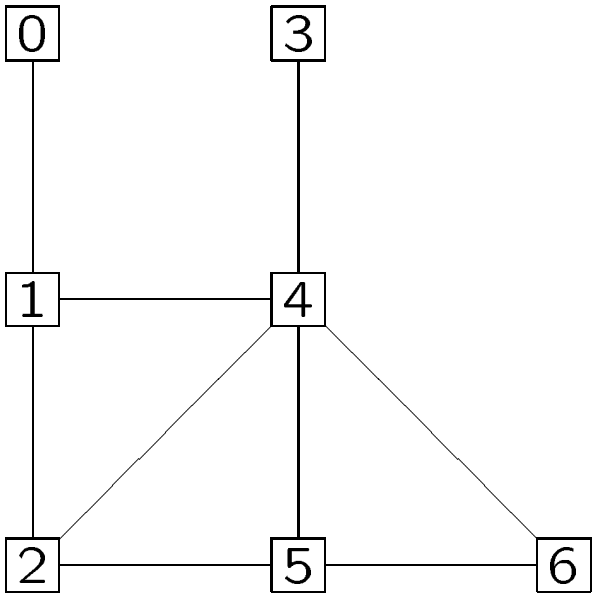
```
public interface Graph {
  int getNumVertices();
  boolean isAdjacent(int x, int y);

  void addEdge(int x, int y);
  void deleteEdge(int x, int y);
}
```

If a graph has v vertices, then we will always assume that the vertex set is $\{0, 1, \ldots, v-1\}$.

All the graphs that we consider will implement the `Graph` interface. As we study more types of graph we will subclass this interface to provide further interfaces.

An example graph

This is a graph X with 7 vertices and 9 edges. We will use it later to illustrate some of the main graph theoretic concepts.



Basic properties of graphs

Let us consider some of the basic terminology of graphs:

Adjacency If $\{x, y\} \in E(G)$, we say that x and y are *adjacent* to each other, and sometimes write $x \sim y$. The number of vertices adjacent to v is called the *degree* or *valency* of v. The vertices adjacent to v are the *neighbours* of v.

Theorem The sum of the degrees of the vertices of a graph is even.

Paths A *path* of *length* n in a graph is a sequence of vertices $v_1 \sim v_2 \sim \cdots \sim v_n+1$ that contains no vertex twice. A *cycle* of length n is a sequence of vertices $v_1 \sim v_2 \sim \cdots v_n \sim v_1$ such that the vertices $\{v_1, v_2, \ldots, v_n\}$ are distinct.

Distance The *distance* between two vertices x and y in a graph is the length of the shortest path between them.

Subgraphs

If G is a graph, then a subgraph H is a graph such that

$$V(H) \subseteq V(G)$$

and

$$E(H) \subseteq E(G)$$

A *spanning* subgraph H has the property that $V(H) = V(G)$—in other words H has been obtained from G only by removing edges.

An *induced* subgraph H must contain every edge of G whose endpoints lie in $V(H)$ — in other words H has been obtained from G by removing vertices and their adjoining edges.

Connectivity, forests and trees

Connected A graph G is *connected* if there is a path between any two vertices. If the graph is not connected then its connected components are the maximal induced subgraphs that are connected.

Forests A forest is a graph that has no cycles.

Trees A tree is a forest with only one connected component. It is easy to see that a tree with n vertices has exactly $n - 1$ edges.

The vertices of degree 1 in a tree are called the *leaves* of the tree.

Directed graphs

There are two important extensions to the basic definition of a graph.

Directed graphs In a directed graph, an edge is an ordered pair of vertices, and hence has a direction. In directed graphs, edges are often called *arcs*.

A directed graph can be used to represent things like a 1-way road system, where travel is possible from x to y, but not from y to x.

We do not need to alter our `Graph` interface at all to represent directed graphs: in fact we will treat all graphs as directed graphs, where an undirected edge between x and y is considered equivalent to an arc from x to y *and* an arc from y to x.

Do we need an interface for `UndirectedGraph`?

```
public interface UndirectedGraph extends Graph {
}
```

Weighted graphs

Weighted graphs In a weighted graph, each of the edges is assigned a weight (usually a non-negative integer). More formally we say that a weighted graph is a graph G together with a weight function $w : E(G) \to \mathbf{R}$ (then $w(e)$ represents the weight of the edge e).

Weights on edges are often used to represent things like travel costs, transmission costs, road capacities and so on. For simplicity we will assume that edge weights are always *integers*. It would be straightforward to implement floating point edge weights.

```
public interface WeightedGraph extends Graph {
  int getWeight(int v, int w);
  void setWeight(int v, int w, int wt);
}
```

We must be careful here about what the effect of `removeEdge()` should be on a weighted graph. We will use the convention that an edge weight of 0 always represents non-adjacency, so that `removeEdge(x,y)` can be implemented as `setWeight(x,y,0)`.

Distance in weighted graphs

When talking about weighted graphs, we need to extend the concept of distance.

Definition In a weighted graph X a path

$$x = x_0 \sim x_1 \sim \cdots \sim x_n = y$$

has *weight*

$$\sum_{i=0}^{i=n-1} w(x_i, x_{i+1}).$$

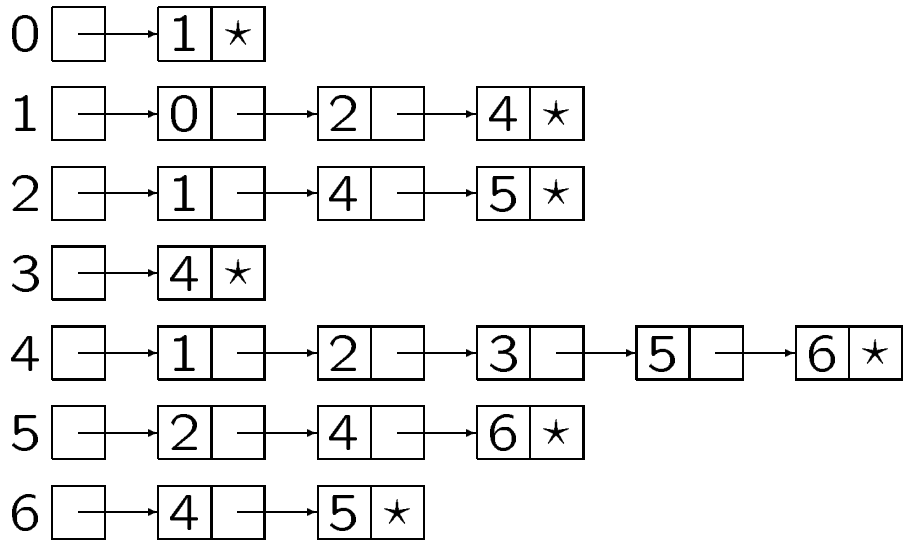
The *shortest path* between two vertices x and y is the path of minimum weight (this should really be called the *lightest path*!).

If we view an unweighted graph as being a special kind of weighted graph, with all the edges of weight 1, then the definition of path reduces to our previous definition.

Representation of graphs

There are two main ways to represent a graph — adjacency lists or an adjacency matrix.

Adjacency lists The graph G is represented by an array of $|V(G)|$ linked lists, with each list containing the neighbours of a vertex.



This representation requires two list elements for each edge and therefore the space required is $\Theta(|V(G)| + |E(G)|)$.

NOTE: In general to avoid writing $|V(G)|$ and $|E(G)|$ we shall simply put $V = |V(G)|$ and $E = |E(G)|$.

This representation is immediately suitable for directed graphs, and requires only minor modification for weighted graphs.

Adjacency matrix

The *adjacency matrix* of a graph G is a $V \times V$ matrix A where the rows and columns are indexed by the vertices and such that $A_{ij} = 1$ if and only if vertex i is adjacent to vertex j.

For the graph X we have the following

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

The adjacency matrix representation uses $\Theta(V^2)$ space.

For a *sparse* graph E is much less than V^2, and hence we would normally prefer the adjacency list representation.

For a *dense* graph E is close to V^2 and the adjacency matrix representation is preferred.

Breadth-first search

Searching through a graph is one of the most fundamental of all algorithmic tasks, and therefore we shall examine several techniques for doing so.

Breadth-first search is a simple but extremely important technique for searching a graph. This search technique starts from a given source vertex s and constructs a spanning tree T for G, called the *breadth-first tree*. It uses a (first-in, first-out) *queue* as its main data structure.

Following CLR, as the search progresses, we will divide the vertices of the graph into three categories, *black* vertices which are the vertices that have been fully examined and incorporated into the tree, *grey* vertices which are the vertices that have been seen (because they are adjacent to a tree vertex) and placed on the queue, and *white* vertices, which have not yet been examined.

Breadth-first search initialization

The final breadth-first tree will be stored as an array called π where $\pi(x)$ is the immediate parent of x in the spanning tree. Of course, as s is the root of this tree, $\pi(s)$ will remain undefined.

To initialize the search we mark the colour of every vertex as *white* and the queue is empty. Then the first step is to mark the colour of v to be *grey*, put $\pi(s)$ to be undefined and add s to the queue.

Breadth-first search repetitive step

Then the following procedure is repeated until the queue is empty.

Take vertex w from the head of the queue
for each vertex x adjacent to w **do**
    if x is white **then**
        $\pi(x) = w$
        Colour x *grey*.
        Add x to the queue.
    end if
end for
Colour w black.

At the end of the search, every vertex in the graph will have colour *black* and the parent or predecessor array π will contain the details of the breadth-first search tree.

Queues revisited

Recall that a *queue* is a data structure whereby the element taken off the data structure is the element that has been on the queue for the longest time.

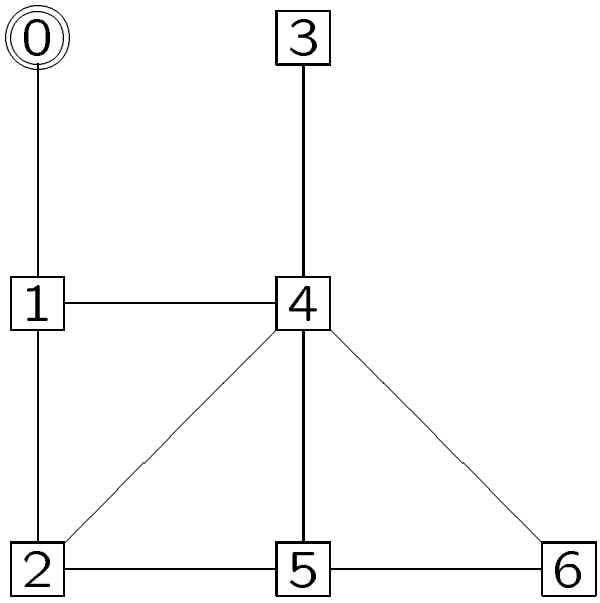
If the maximum length of the queue is known in advance (and is not too great) then a queue can be very efficiently implemented by simply using an array.

An array of n elements is initialized, and two pointers called *head* and *tail* are maintained — the head gives the location of the next element to be removed, while the tail gives the location of the first empty space in the array.

It is trivial to see that both enqueueing and dequeing operations take $\Theta(1)$ time.

See CLR, Section 11.1 for further details.

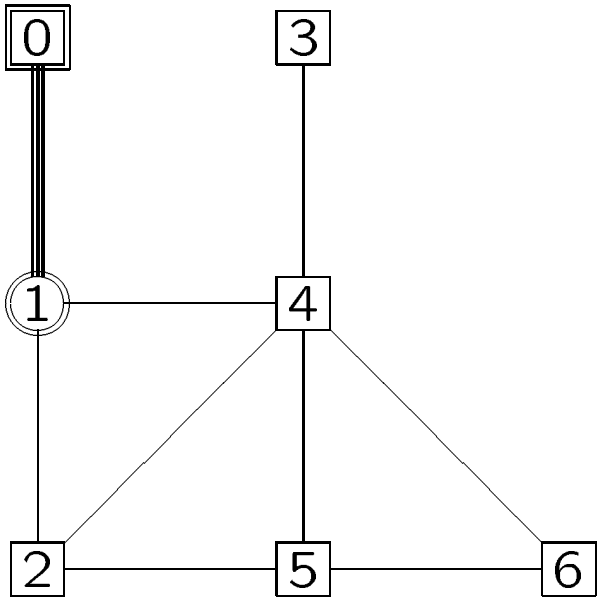
Example of breadth-first search



Head
↓
queue 0

x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*white*	
2	*white*	
3	*white*	
4	*white*	
5	*white*	
6	*white*	

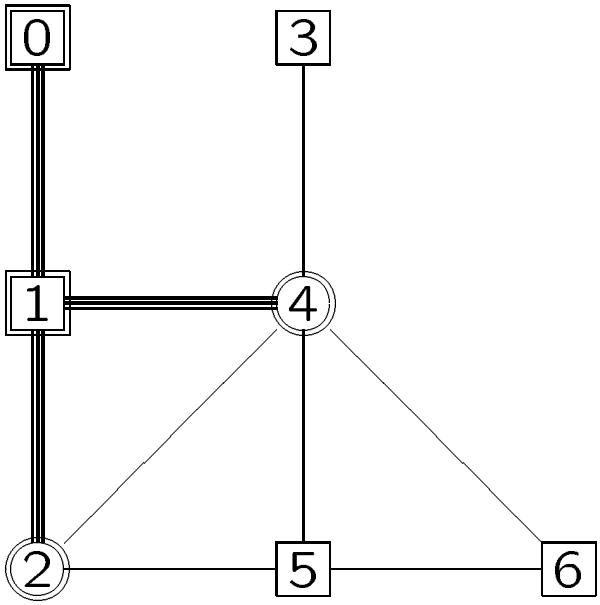
After visiting vertex 0



Head
↓
queue 0 1

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*grey*	0
2	*white*	
3	*white*	
4	*white*	
5	*white*	
6	*white*	

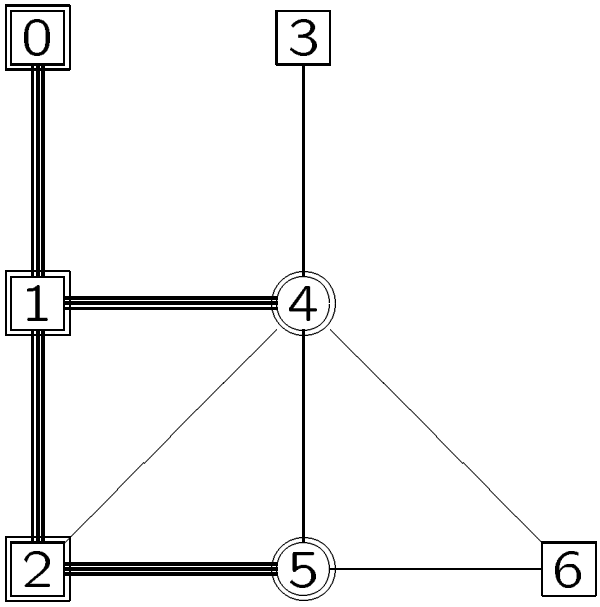
After visiting vertex 1



Head
↓
queue 0 1 2 4

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*grey*	1
3	*white*	
4	*grey*	1
5	*white*	
6	*white*	

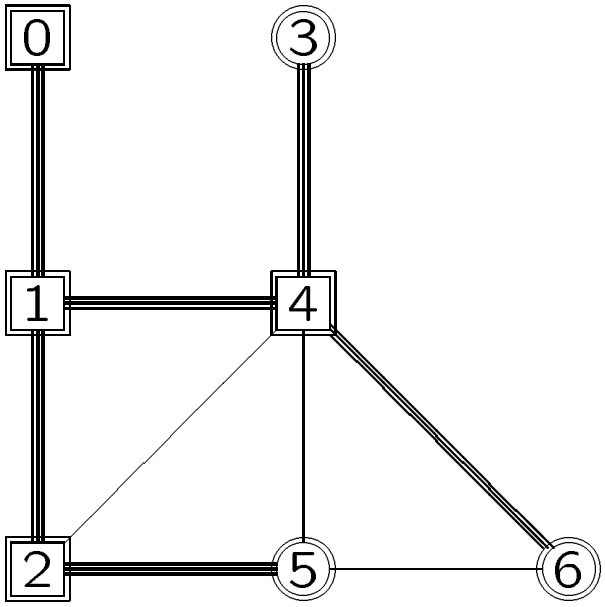
After visiting vertex 2



Head
↓
queue 0 1 2 4 5

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*black*	1
3	*white*	
4	*grey*	1
5	*grey*	2
6	*white*	

After visiting vertex 4

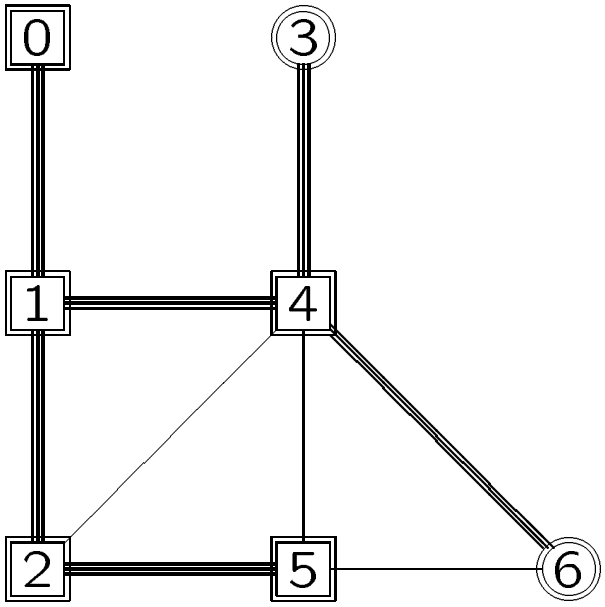


Head
↓

queue 0 1 2 4 5 3 6

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*black*	1
3	*grey*	4
4	*black*	1
5	*grey*	2
6	*grey*	4

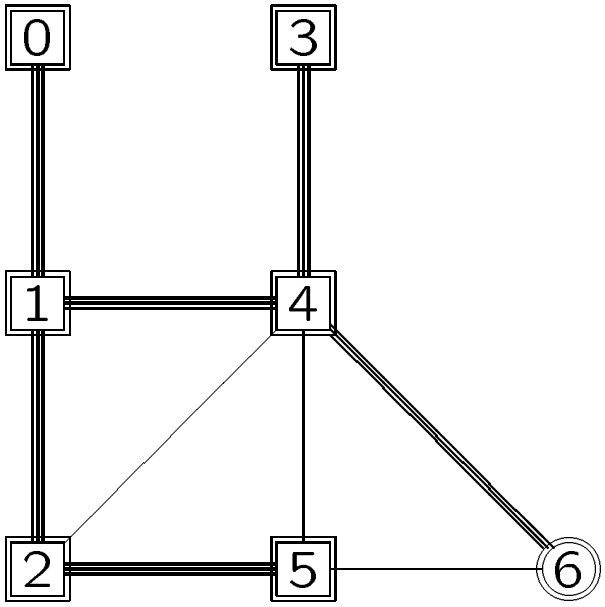
After visiting vertex 5



Head
↓
queue 0 1 2 4 5 3 6

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*black*	1
3	*grey*	4
4	*black*	1
5	*black*	2
6	*grey*	4

After visiting vertex 3

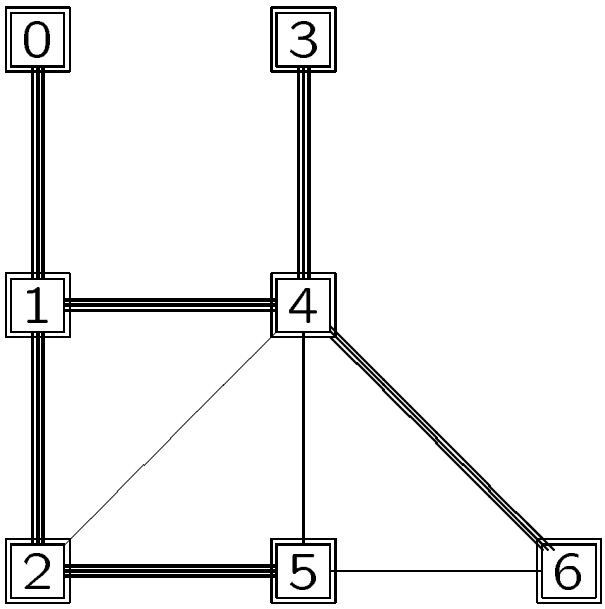


Head
↓

queue 0 1 2 4 5 3 6

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*black*	1
3	*black*	4
4	*black*	1
5	*black*	2
6	*grey*	4

After visiting vertex 6



Head
↓

queue 0 1 2 4 5 3 6

x	colour(x)	$\pi(x)$
0	*black*	**undef**
1	*black*	0
2	*black*	1
3	*black*	4
4	*black*	1
5	*black*	2
6	*black*	4

At termination

At the termination of breadth-first search, every vertex in the same connected component as s is a black vertex and the array π contains details of a spanning tree for that component—the breadth-first search tree.

Complexity analysis

During the breadth-first search each vertex is enqueued once and dequeued once. As each enqueueing/dequeuing operation takes constant time, the queue manipulation takes $\Theta(V)$ time. At the time the vertex is dequeued, the adjacency list of that vertex is completely examined. Therefore we take $\Theta(E)$ time examining all the adjacency lists and the total time is $\Theta(V + E)$.

Uses of BFS

Breadth-first search is particularly useful for certain simple tasks such as determining whether a graph is connected, or finding the distance between two vertices.

The vertices of G are examined in order of increasing distance from s—first s, then its neighbours, then the vertices at distance 2 from s and so on. The spanning tree constructed provides a shortest path from any vertex back to s just by following the array π.

Therefore it is simple to modify the breadth-first search to provide an array of distances d where $d(v)$ is the distance of the vertex v from the source vertex s.

Breadth-first search finding distances

To initialize the search we mark the colour of every vertex as *white* and the queue is empty. Then the first step is to colour the source vertex s *grey*, put $\pi(s)$ to be undefined, $d(s) = 0$, and add s to the queue.

Then until the queue is empty we repeat the following procedure.

Take vertex w from the head of the queue
for each vertex x adjacent to w **do**
    if x is *white* **then**
        $d(x) = d(w) + 1$
        $\pi(x) = w$
        Colour x *grey*.
        Add x to the queue.
    end if
end for
Colour w *black*.

Depth-first search

Depth-first search is another important technique for searching a graph. Similarly to breadth-first search it also computes a spanning tree for the graph, but the tree is very different.

The structure of depth-first search is naturally *recursive* so we will give a recursive description of it. Nevertheless it is useful and important to consider the non-recursive implementation of the search.

The fundamental idea behind depth-first search is to visit the next unvisited vertex, thus extending the current path as far as possible. When the search gets stuck in a "corner" we back up along the path until a new avenue presents itself (this is called *backtracking*).

Basic recursive depth-first search

The following recursive program computes the depth-first search tree for a graph G starting from the source vertex s.

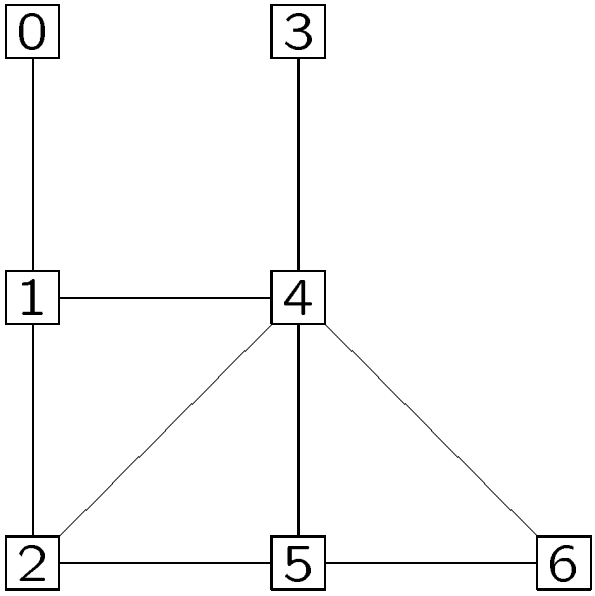
To initialize the search we mark the colour of every vertex as *white*. Then we call the recursive routine DFS(s) where s is the source vertex.

procedureDFS(w)
Colour w *grey*.
for each vertex x adjacent to w **do**
    if x is *white* **then**
        $\pi(x) = w$
        DFS(x)
    end if
end for
Colour w *black*.

At the end of this depth-first search procedure we have produced a spanning tree containing every vertex in the connected component containing s.

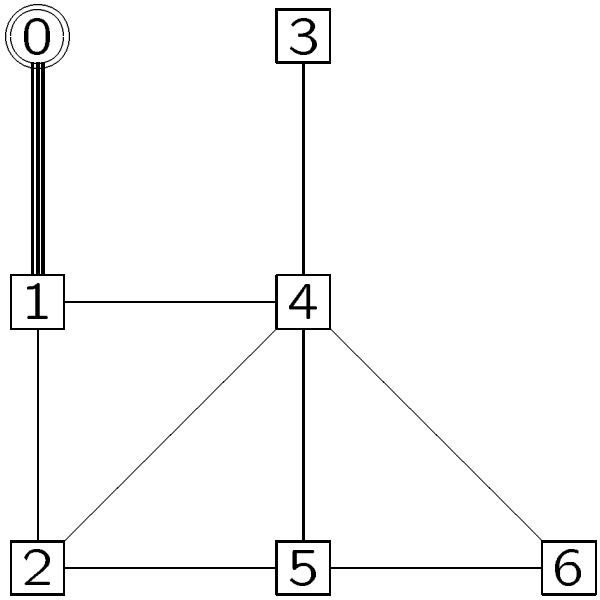
Example of depth-first search

We will search the following graph from the source vertex 0.



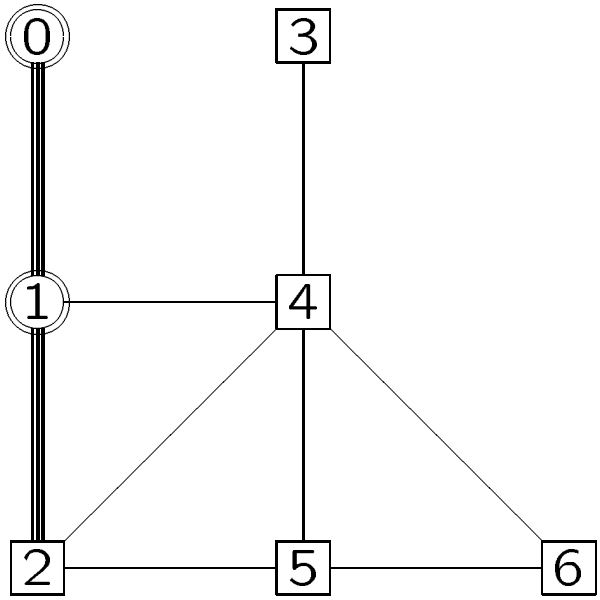
x	colour(x)	$\pi(x)$
0	*white*	**undef**
1	*white*	
2	*white*	
3	*white*	
4	*white*	
5	*white*	
6	*white*	

Immediately prior to calling DFS(1)



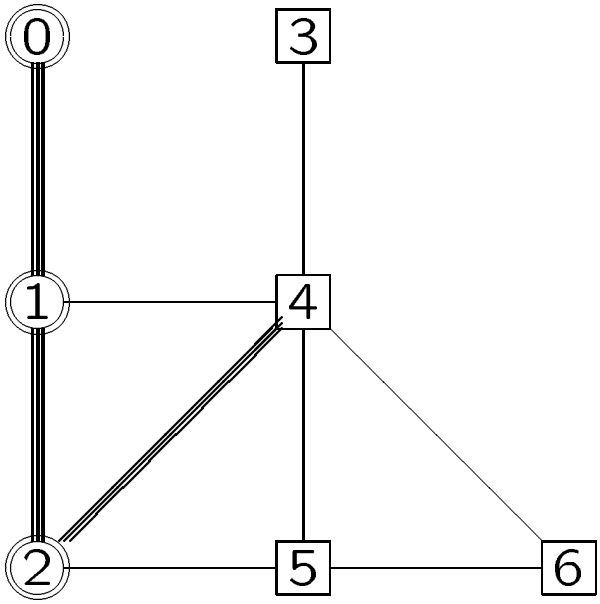
x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*white*	0
2	*white*	
3	*white*	
4	*white*	
5	*white*	
6	*white*	

Immediately prior to calling DFS(2)



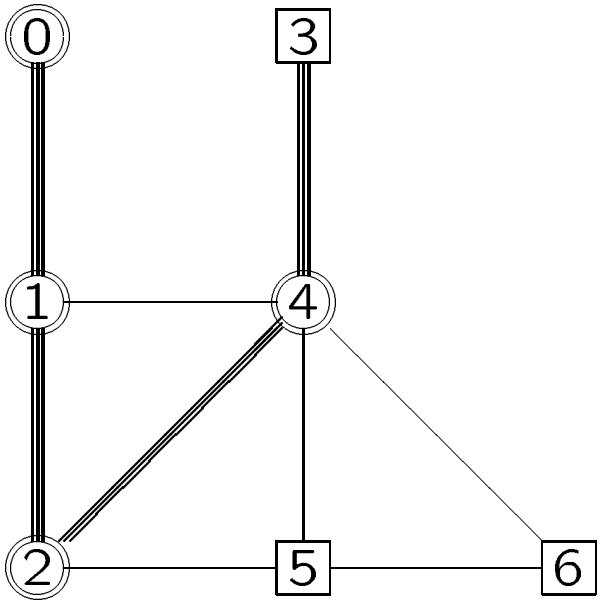
x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*grey*	0
2	*white*	1
3	*white*	
4	*white*	
5	*white*	
6	*white*	

Immediately prior to calling DFS(4)



x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*grey*	0
2	*grey*	1
3	*white*	
4	*white*	2
5	*white*	
6	*white*	

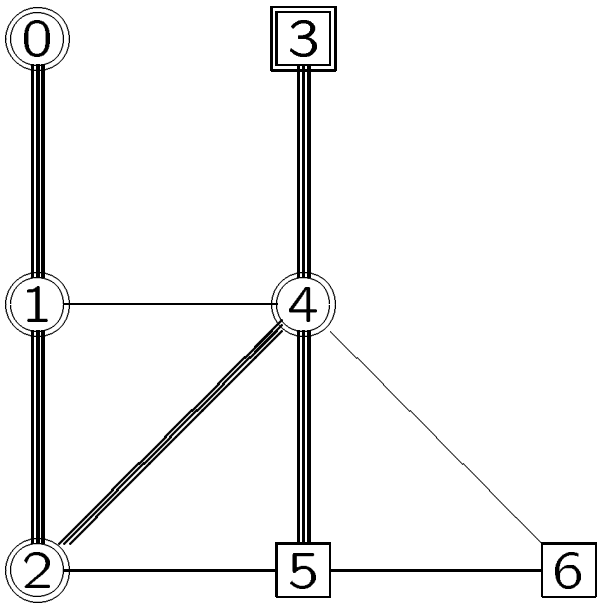
Immediately prior to calling DFS(3)



x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*grey*	0
2	*grey*	1
3	*white*	4
4	*grey*	2
5	*white*	
6	*white*	

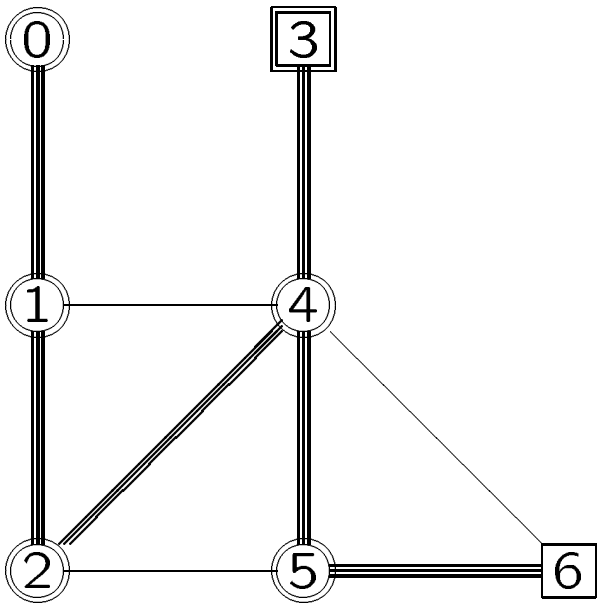
Immediately prior to calling DFS(5)

Now the call to DFS(2) actually finishes without making any more recursive calls so we return to examining the neighbours of vertex 4, the next of which is vertex 5.



x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*grey*	0
2	*grey*	1
3	*black*	4
4	*grey*	2
5	*white*	4
6	*white*	

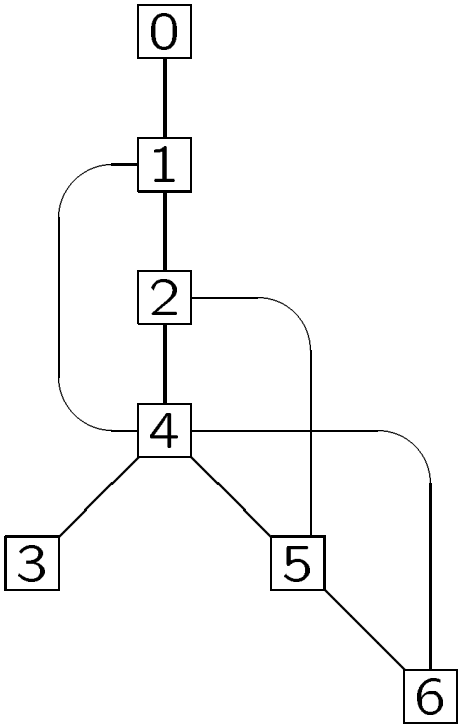
Immediately prior to calling DFS(6)



x	colour(x)	$\pi(x)$
0	*grey*	**undef**
1	*grey*	0
2	*grey*	1
3	*black*	4
4	*grey*	2
5	*grey*	4
6	*white*	5

The depth-first search tree

After completion of the search we can draw the depth-first search tree for this graph:



In this picture the slightly thicker straight edges are the *tree edges* and the remaining edges are the *back edges*—the back edges arise when we examine an edge (u, v) and discover that its endpoint v no longer has the colour *white*

Analysis of DFS

The running time of DFS is easy to analyse as follows.

First we observe that the routine DFS(w) is called exactly once for each vertex w; during the execution of this routine we perform only constant time array accesses, and run through the adjacency list of w once.

Running through the adjacency list of each vertex exactly once takes $\Theta(E)$ time overall, and hence the total time taken is $\Theta(V + E)$.

In fact, we can say more and observe that because every vertex and every edge are examined precisely once in both BFS and DFS, the time taken is $\Theta(V + E)$.

Discovery and finish times

The operation of depth-first search actually gives us more information than simply the depth-first search tree; we can assign two times to each vertex.

Consider the following modification of the search, where *time* is a global variable that starts at time 1.

procedure DFS(w)
colour[w] $\leftarrow$ *grey*
discovery[w] $\leftarrow$ *time*
time $\leftarrow$ *time+1*
for each vertex x adjacent to w **do**
    if *colour[x]* is *white* **then**
        $\pi[x] \leftarrow w$
        DFS(x)
    end if
end for
colour[w] $\leftarrow$ *black*
finish[w] $\leftarrow$ *time*
time $\leftarrow$ *time+1*

The parenthesis property

This assigns to each vertex a *discovery* time, which is the time at which it is first discovered, and a *finish* time, which is the time at which all its neighbours have been searched and it no longer plays any further role in the search.

The discovery and finish times satisfy a property called the *parenthesis property*.

Imagine writing down an expression consisting entirely of labelled parentheses — at the time of discovering vertex u we open a parenthesis $(_u$ and a the time of finishing with u we close the parenthesis $_u)$.

Then the resulting expression is a well-formed expression with correctly nested parentheses.

For our example depth-first search we get:

$$(_0\ (_1\ (_2\ (_4\ (_3\ _3)\ (_6\ (_6\ _6)\ _5)\ _4)\ _2)\ _1)\ _0)$$

Depth-first search for directed graphs

A depth-first search on an undirected graph produces a classification of the edges of the graph into *tree edges*, or *back edges*. For a directed graph, there are further possibilities. The same depth-first search algorithm can be used to classify the edges into four types:

tree edges If the procedure DFS(u) calls DFS(v) then (u, v) is a tree edge

back edges If the procedure DFS(u) explores the edge (u, v) but finds that v is an already visited ancestor of u, then (u, v) is a back edge

forward edges If the procedure DFS(u) explores the edge (u, v) but finds that v is an already visited descendant of u, then (u, v) is a forward edge

cross edges All other edges are cross-edges

Topological sort

We shall consider a classic simple application of depth-first search.

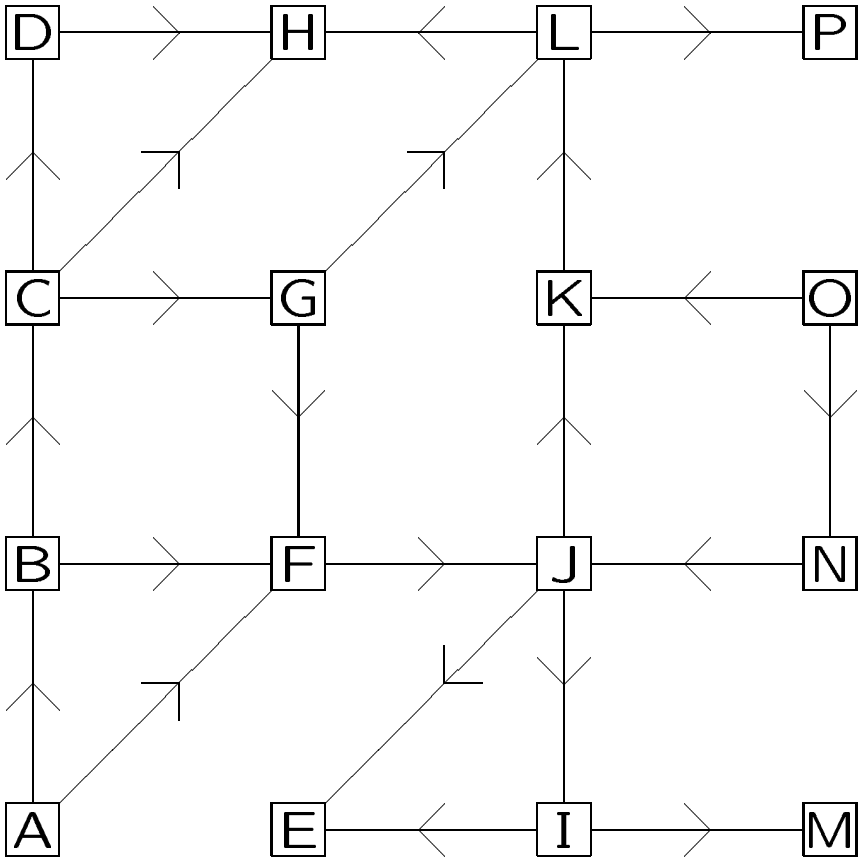
Definition A *directed acyclic graph (dag)* is a directed graph with no directed cycles.

Theorem In a depth-first search of a dag there are no back edges.

Consider now some complicated process in which various jobs must be completed before others are started. We can model this by a graph D where the vertices are the jobs to be completed and there is an edge from job u to job v if job u must be completed before job v is started. Our aim is to find some linear ordering of the jobs such that they can be completed without violating any of the constraints.

This is called finding a *topological sort* of the dag D.

Example of a dag to be topologically sorted



What is the appropriate linear order in which to do these jobs so that all the precedences are satisfied.

Algorithm for TOPOLOGICAL SORT

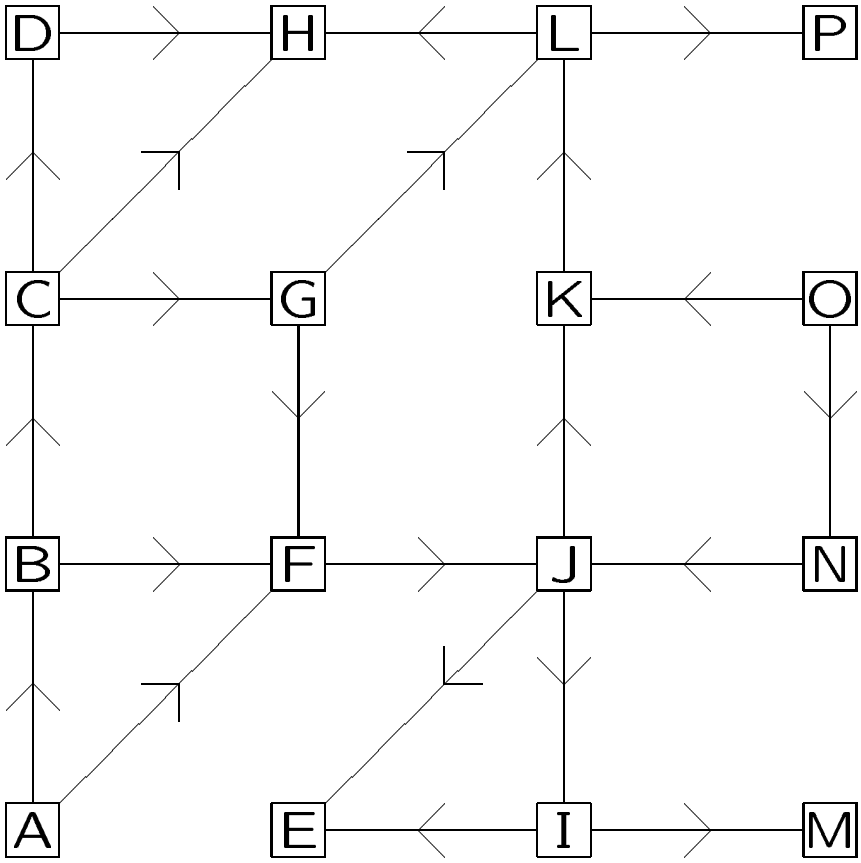
The algorithm for topological sort is an extremely simple application of depth-first search.

Algorithm

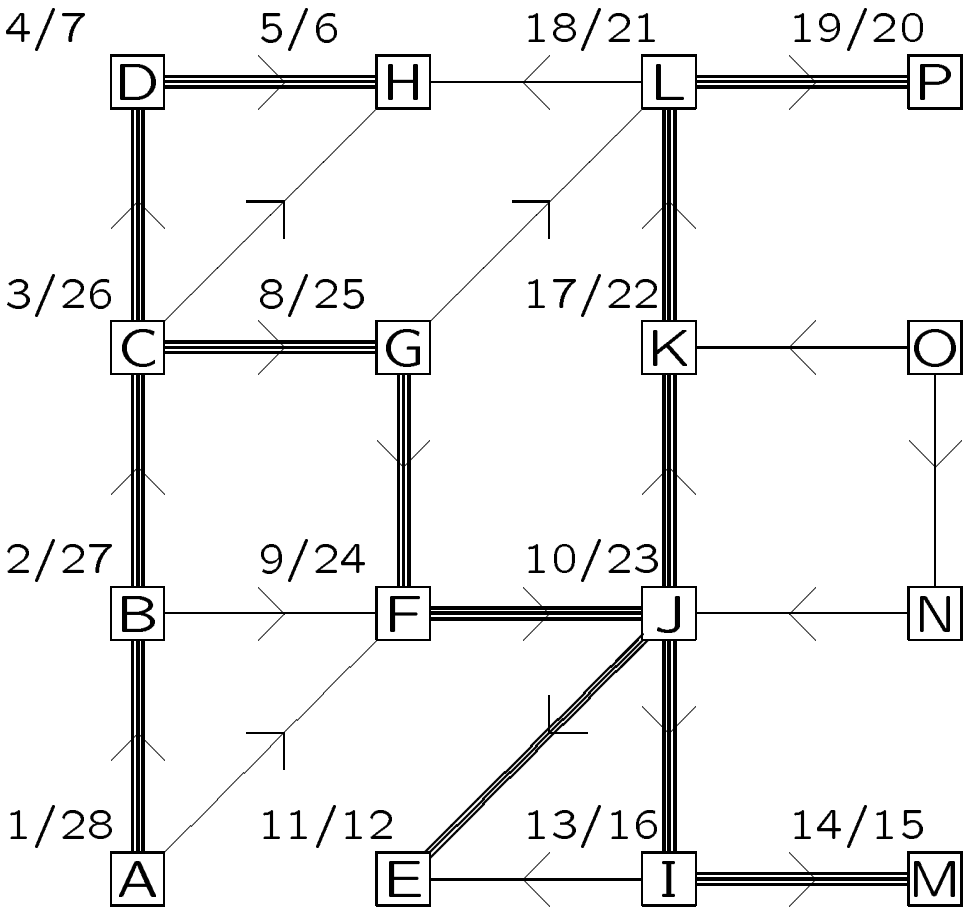
Apply the depth-first search procedure to find the finishing times of each vertex. As each vertex is finished, put it onto the *front* of a linked list.

At the end of the depth-first search the linked list will contain the vertices in topologically sorted order.

Doing the topological sort

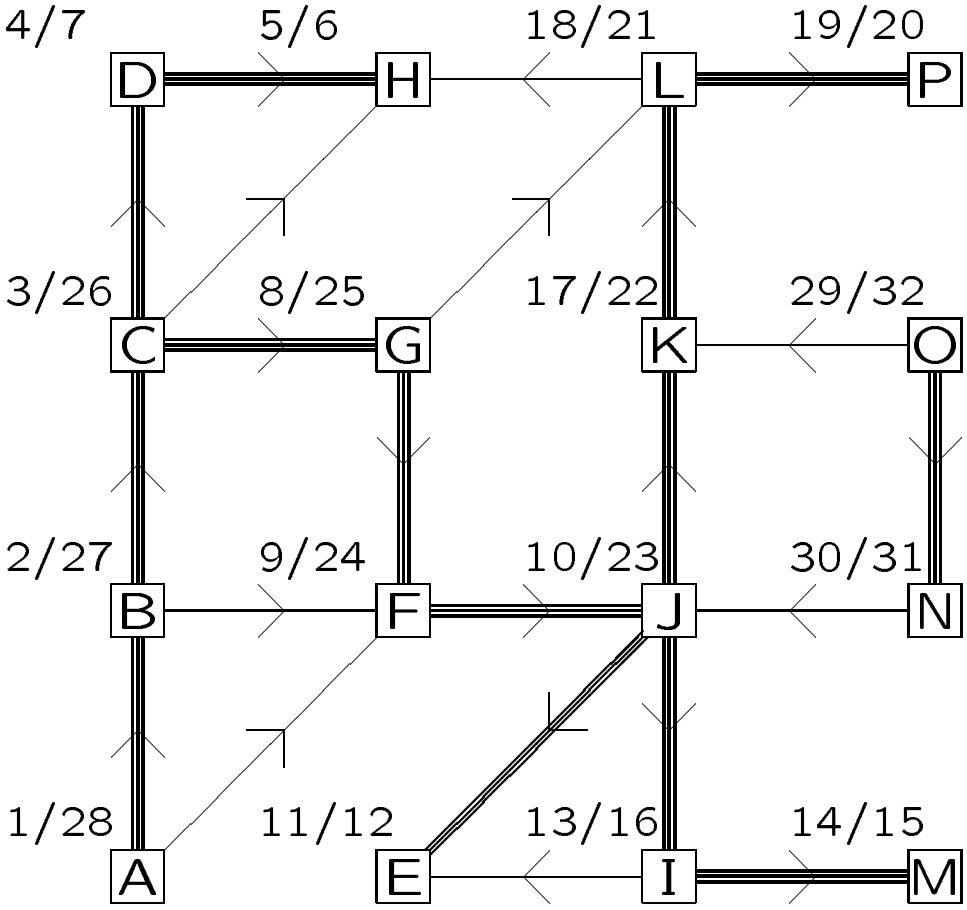


After the first depth-first search



Notice that there is a component that has not been reached by the depth-first search. To complete the search we just repeatedly perform depth-first searches until all vertices have been examined.

After the entire search



As the vertices were placed at the front of a linked list as they became finished the final topological sort is: $O - N - A - B - C - G - F - J - K - L - P - I - M - E - D - H$

A topologically sorted dag has the property that any edges drawn in the above diagram will got from left-to-right.

Analysis and correctness

Time analysis of the algorithm is very easy — to the $\Theta(V + E)$ time for the depth-first search we must add $\Theta(V)$ time for the manipulation of the linked list. Therefore the total time taken is again $\Theta(V + E)$.

Why does it work?

We shall try to show that for any edge (u, v) in the dag the finishing time $f(u) > f(v)$.
Consider the stage at which the edge (u, v) is encountered. If (u, v) is a tree edge, then the depth-first search proceeds from u to v and clearly finishes with v before finally returning to u. On the other hand if (u, v) is a forward or cross edge, then the vertex v has already been completely examined by this stage, and hence vertex u must have a later finishing time. The edge (u, v) cannot be a back edge because a dag has no cycles.

Other uses for DFS

DFS is the standard algorithmic method for solving the following two problems:

Strongly connected components In a directed graph D a strongly connected component is a maximal subset S of the vertices such that for any two vertices u, $v \in S$ there is a directed path from u to v and from v to u.

Depth-first search can be used on a digraph to find strongly connected components in time $\Theta(V + E)$.

Biconnected components In a connected graph G, an *articulation point* is a vertex whose removal disconnects the graph.

Depth-first search can be used on a graph to find all the articulation points in time $\Theta(V + E)$.

Minimum spanning tree (MST)

Consider a group of villages in a remote area that are to be connected by telephone lines. There is a certain cost associated with laying the lines between any pair of villages, depending on their distance apart, the terrain and some pairs just cannot be connected.

Our task is to find the minimum possible cost in laying lines that will connect all the villages.

This situation can be modelled by a weighted graph W, in which the weight on each edge is the cost of laying that line. A *minimum spanning tree* in a graph is a subgraph that is (1) a spanning subgraph (2) a tree and (3) has a lower weight than any other spanning tree.

It is clear that finding a MST for W is the solution to this problem.

The greedy method

Definition A *greedy algorithm* is an algorithm in which at each stage a locally optimal choice is made.

A greedy algorithm is therefore one in which no overall strategy is followed, but you simply do whatever looks best at the moment.

For example a mountain climber using the greedy strategy to climb Everest would at every step climb in the steepest direction. From this analogy we get the computational search technique known as *hill-climbing*.

In general greedy methods have limited use, but fortunately, the problem of finding a minimum spanning tree can be solved by a greedy method.

Kruskal's method

Kruskal invented the following very simple method for building a minimum spanning tree. It is based on building a forest of lowest possible weight and continuing to add edges until it becomes a spanning tree.

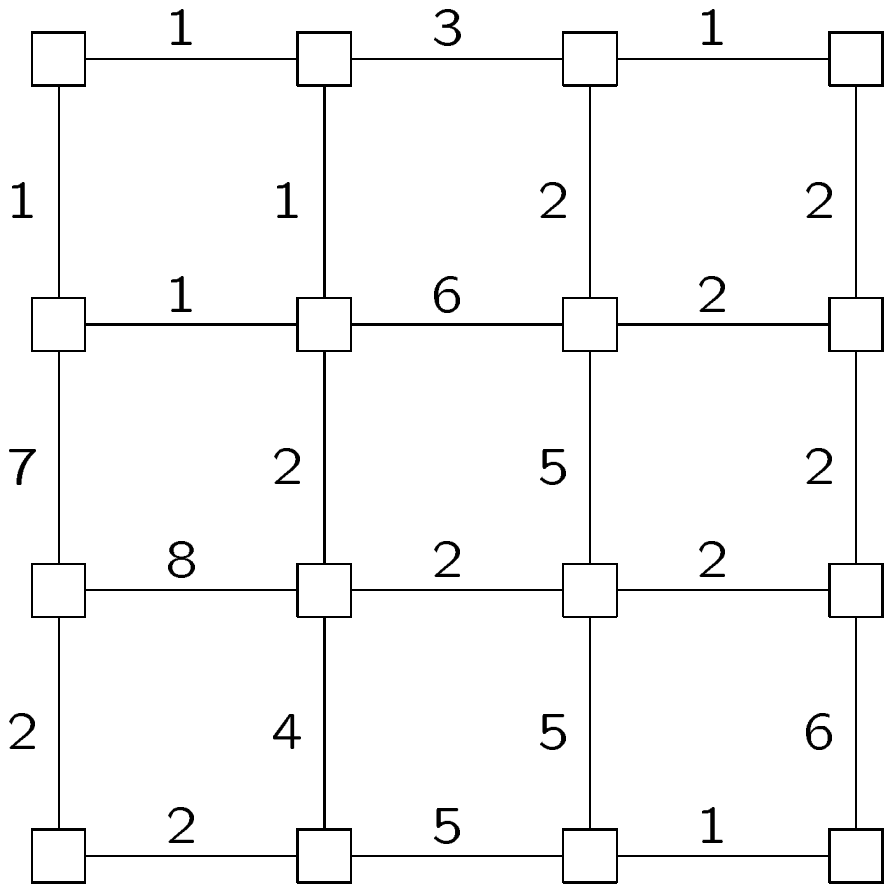
Kruskal's method

Initialize F to be the forest with all the vertices of G but none of the edges.

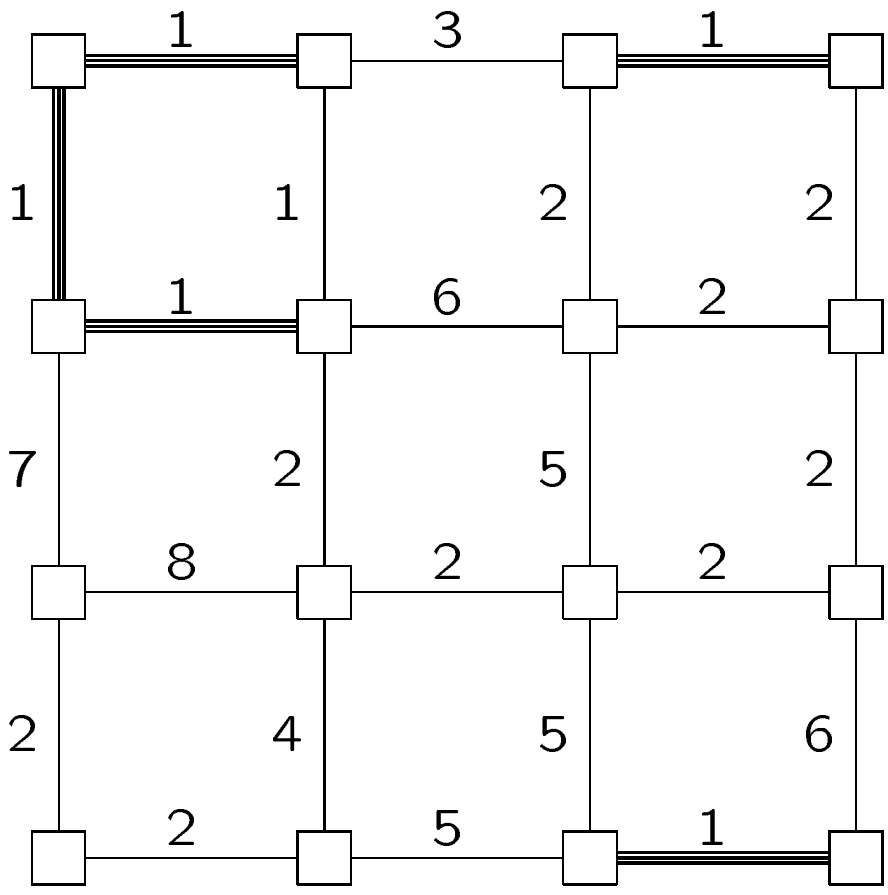
repeat
Pick an edge e of minimum possible weight
if $F \cup \{e\}$ is a forest **then**
$F \leftarrow F \cup \{e\}$
end if
until F contains $n - 1$ edges

Therefore we just keep on picking the smallest possible edge, and adding it to the forest, providing that we never create a cycle along the way.

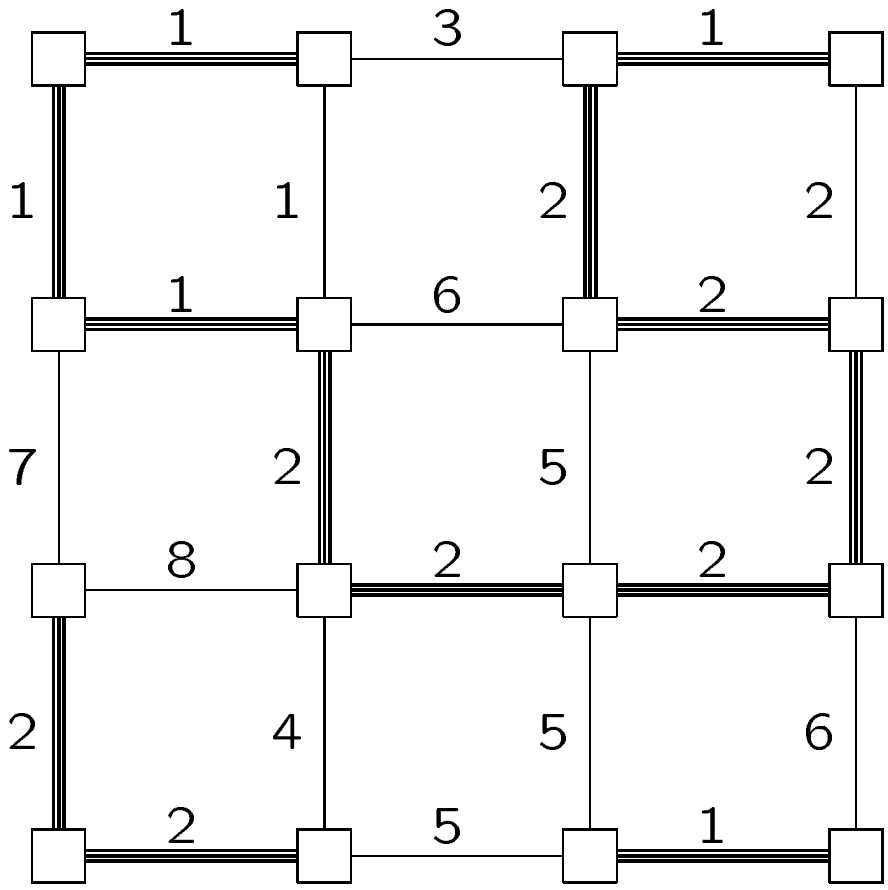
Example



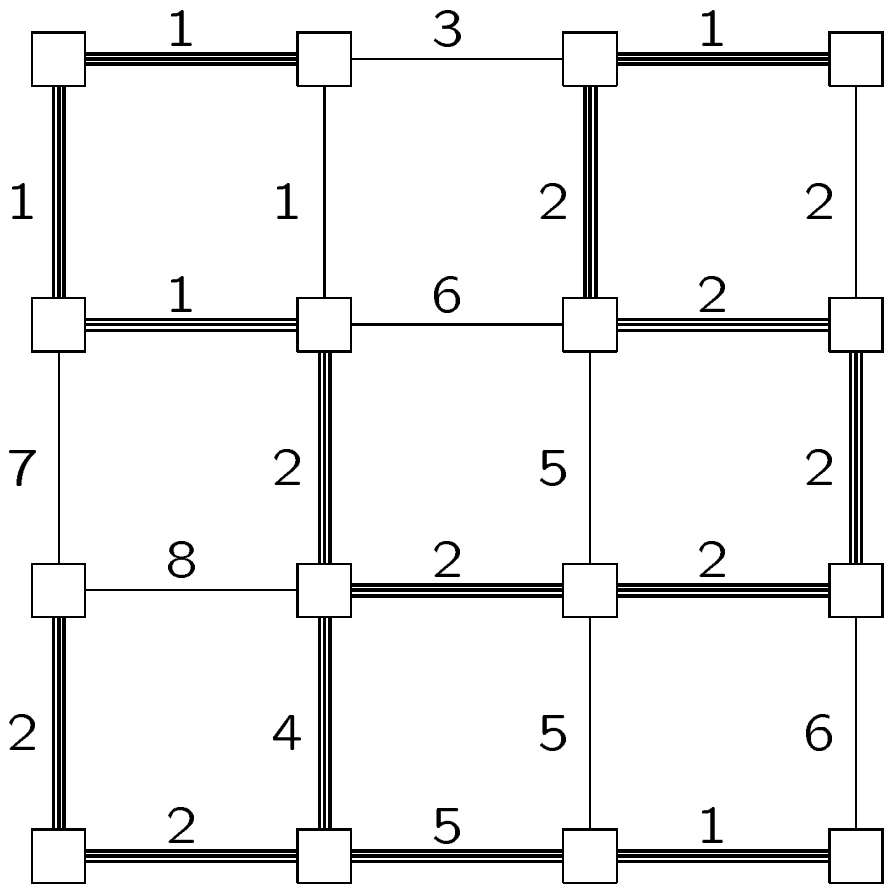
After using edges of weight 1



After using edges of weight 2



The final MST



Prim's algorithm

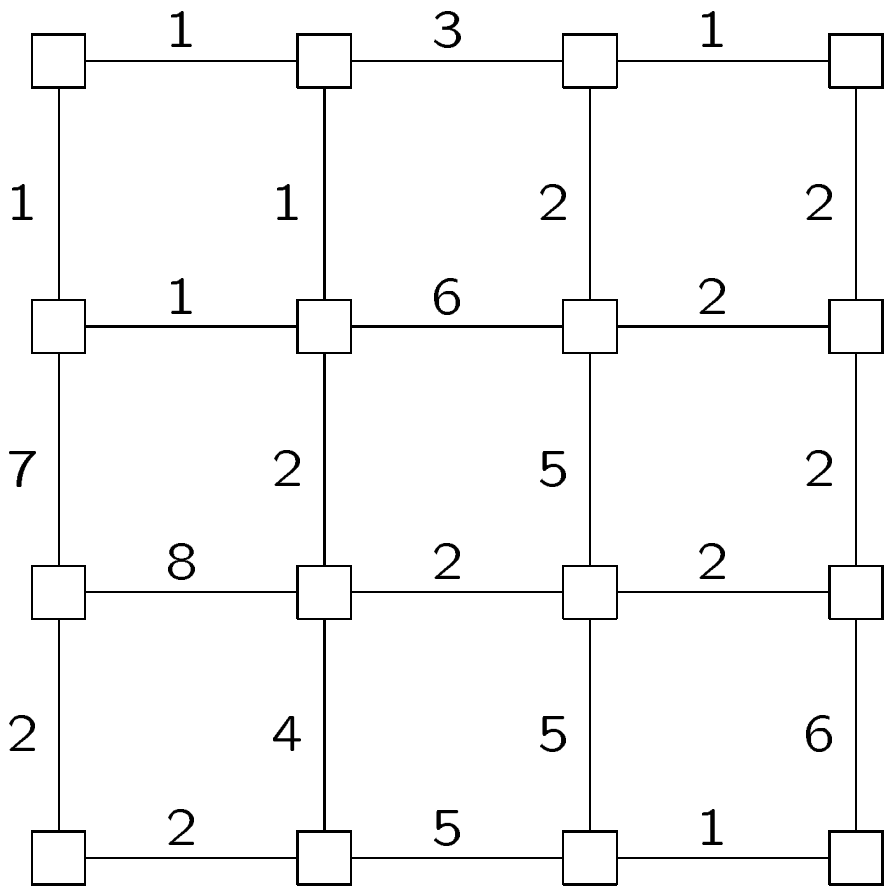
Prim's algorithm is another greedy algorithm for finding a minimum spanning tree.

The idea behind Prim's algorithm is to grow a minimum spanning tree edge-by-edge by always adding the shortest edge that touches a vertex in the current tree.

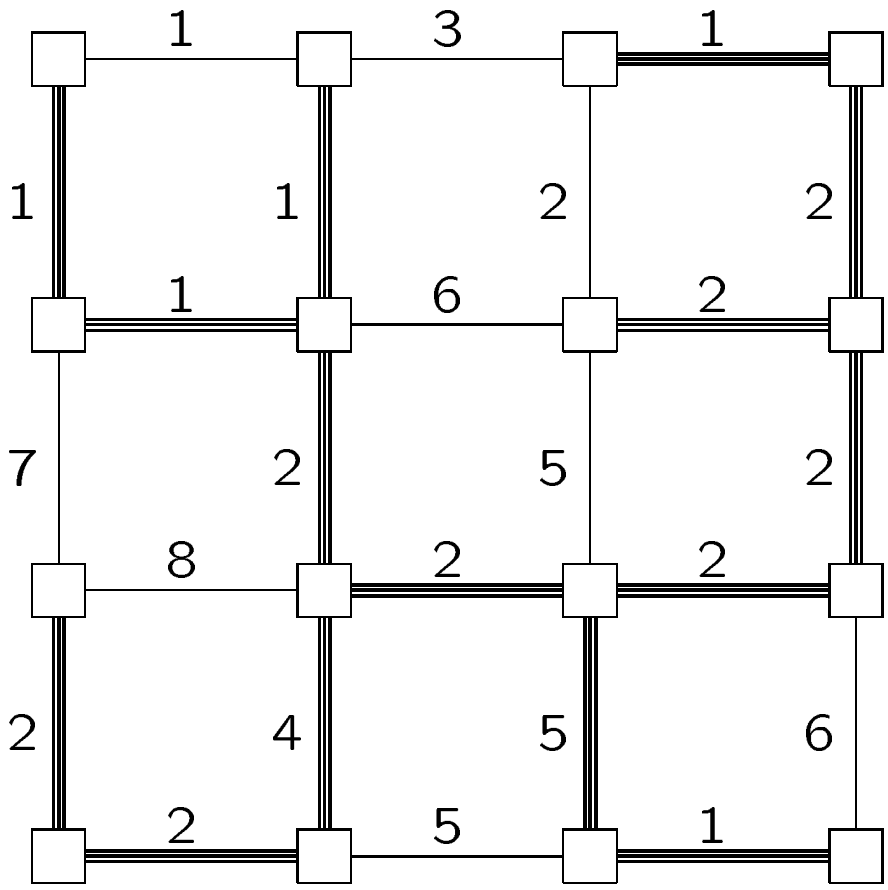
Notice the difference between the algorithms: Kruskal's algorithm always maintains a spanning subgraph which only becomes a tree at the final stage.

On the other hand, Prim's algorithm always maintains a tree which only becomes spanning at the final stage.

Prim's algorithm in action



One solution



Problem solved?

As far as a mathematician is concerned the problem of a minimum spanning tree is well-solved. We have two simple algorithms both of which are guaranteed to find the best solution. (After all, a greedy algorithm must be one of the simplest possible).

In fact, the reason why the greedy algorithm works in this case is well understood — the collection of all the subsets of the edges of a graph that do not contain a cycle forms what is called a (graphic) **matroid**.

Loosely speaking, a greedy algorithm always works on a matroid and never works otherwise.

Implementation issues

In fact the problem is far from solved because we have to decide how to *implement* the two greedy algorithms.

The details of the implementation of the two algorithms are interesting because they use (and illustrate) two important data structures — the *partition* and the *priority queue*.

Implementation of Kruskal

The main problem in the implementation of Kruskal is to decide whether the next edge to be added is allowable — that is, does it create a cycle or not.

Suppose that at some stage in the algorithm the next shortest edge is $\{x, y\}$. Then there are two possibilities:

x and y lie in different trees of F: In this case adding the edge does not create any new cycles, but merges together two of the trees of F

x and y lie in the same tree of F: In this case adding the edge creates a cycle and the edge should not be added to F

Therefore we need data structures that allow us to quickly find the tree to which an element belongs and quickly merge two trees.

Union/find data structure

A *partition* of a set Ω is a collection of disjoint sets that cover Ω.

At the beginning of Kruskal's algorithm we have a partition of the vertices into the discrete partition where each cell has size 1. As new edges are added, we need to determine whether its two ends are in the same cell or not, and if not we need to merge the two cells.

Therefore we need an ADT that supports these operations. In Java we express our desire as an interface.

```
public interface UnionFind {
  void union(int x, int y);
  int find(int x);
}
```

The naive solution

One simple way to represent a partition is simply to choose one element of each cell to be the "leader" of that cell. Then we can simply keep a private array π of length n where $\pi(x)$ is the leader of the cell containing x.

Example Consider the partition of 8 elements into 3 cells as follows:

$$\{0, 2 \mid 1, 3, 5 \mid 4, 6, 7\}$$

We could represent this as an array as follows

x	0	1	2	3	4	5	6	7
$\pi(x)$	0	1	0	1	4	1	4	4

Then certainly the method `find()` is straightforward — we can decide whether x and y are in the same cell just by comparing $\pi(x)$ with $\pi(y)$.

Thus `find()` has complexity $\Theta(1)$.

Updating the partition

Suppose now that we wish to update the partition by merging the first two cells to obtain the partition

$$\{0, 1, 2, 3, 5 \mid 4, 6, 7\}$$

We could update the data structure by running through the entire array π and updating it as necessary.

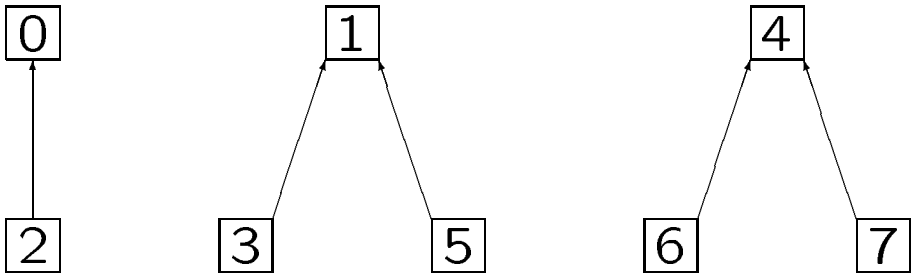
x	0	1	2	3	4	5	6	7
$\pi(x)$	0	0	0	0	4	0	4	4

This takes time $\Theta(n)$, and hence the `union` method is slow.

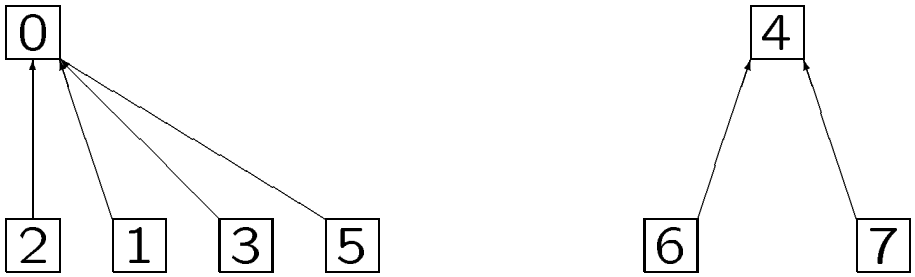
Can we improve the time of this `union` method?

A disjoint sets forest

Consider the following graphical representation of the union/find data structure above, where each element points (upwards) to the "leader" of the cell that it is in.

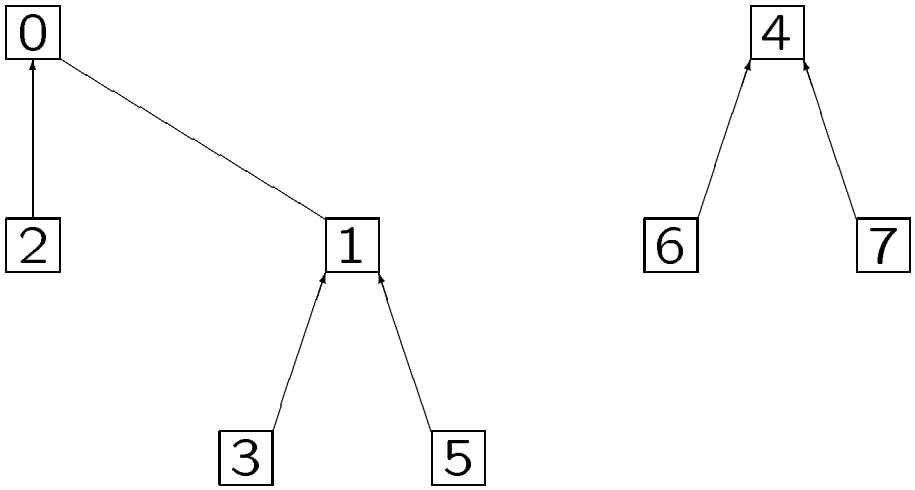


Now mergine two cells is accomplished by adjusting the pointers so they point to the new leader.



However we can achieve something similar by just adjusting one pointer—suppose we simply change the pointer for the element 1, by making it point to 0 instead of itself.

The new data structure



This improved merging has takes only $\Theta(1)$. However we have lost the ability to do `find()` properly. In order to correctly find the leader of the cell containing an element we have to run through a little loop:

```
int find(int x) {

  while (x != pi[x])
    x = pi[x];

  return x;
}
```

Unfortunately, this new `find()` operation may take time $O(n)$ so we seem to have gained nothing.

Union-by-rank heuristic

There are two heuristics that can be applied to the new data structure, that speed things up enormously at the cost of maintaining a little extra data.

Let the *rank* of a root node of a tree be the height of that tree (the maximum distance from a leaf to the root).

The *union-by-rank* heuristic tries to keep the trees *balanced* at all times. When a merging operation needs to be done, the root of the shorter tree is made to point to the root of the taller tree. The resulting tree therefore does not increase its height unless both trees are the same height in which case the height increases by one.

The Partition class

```
public class Partition implements UnionFind {

  private int[] pi;
  private int[] rank;

  Partition(int n) {

    pi = new int[n];
    rank = new int[n];

    for (int i=0;i<n;i++) {
      pi[i] = i;
      rank[i] = 0;
    }
  }

  /** methods union and find
      to be added              **/


}
```

Implementation of union by rank

```
/*****************************************
    Note: Assumes lcell1 and lcell2 are cell
    leaders. Call find() if they are not.
*****************************************/

public void union(int lcell1, int lcell2) {
  if (rank[lcell2] > rank[lcell1]) {
    pi[lcell1] = lcell2;
  }
  else {
    pi[lcell2] = lcell1;

    if (rank[lcell2] == rank[lcell1])
      rank[lcell1]++;
  }
}
```

Notice how the rank is updated only if necessary.

Path compression heuristic

The path compression heuristic is based on the idea that when we perform a `find()` operation we have to follow a path from x to the root of the tree containing x.

After we have done this why do we not simply go back down through this path and make all these elements point *directly* to the root of the tree, rather than in a long chain through each other?

This is reminiscent of our naive algorithm, where we made *every* element point directly to the leader of its cell, but it is much cheaper because we only alter things that we needed to look at anyway.

Implementation of path compression

The path compression heuristic is easily implemented, simply by adjusting the `find()` method so that it updates the `pi[]` entry for every element it touches.

```
public int find(int cell) {
  if (cell != pi[cell])
    pi[cell] = find(pi[cell]);
  return pi[cell];
}
```

Make sure that you understand why this simple recursive method implements the path compression heuristic.

Complexity of Kruskal

In the worst case, we will perform E operations on the partition data structure which has size V. By the complicated argument in CLR we see that the total time for these operations if we use both heuristics is $O(E \lg^* V)$.

However we must add to this the time that is needed to sort the edges — because we have to examine the edges in order of length. This time is $O(E \lg E)$ if we use a sorting technique such as quicksort, and hence the overall complexity of Kruskal's algorithm is $O(E \lg E)$.

Notice that the time taken for this algorithm is dominated by sorting the edges. If there are many edges, in particular if the graph is a complete graph, then this can take a long time.

Implementation of Prim

For Prim's algorithm we repeatedly have to select the next vertex that is *closest* to the tree that we have built so far. Therefore we need some sort of data structure that will enable us to associate a value with each vertex (being the distance to the tree under construction) and rapidly select the vertex with the lowest value.

From our study of Data Structures we know that the appropriate data structure is a *priority queue* and that a priority queue is implemented by using a *heap*.

The priority queue ADT

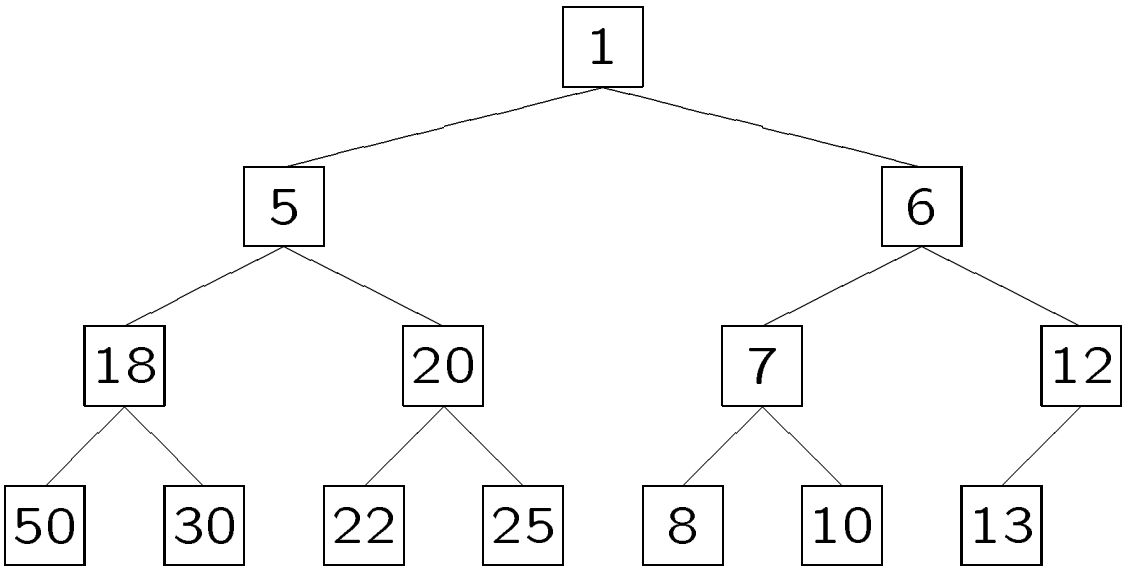
Recall that a *priority queue* is an abstract data type that stores objects with an associated value. We consider these objects to be `Comparable` where the comparisons are performed according to the value associated with the object.

A priority queue allows the objects with lowest values to be examined and extracted from the queue.

```
public interface PriorityQueue {
  void insert(Comparable x);
  Comparable deleteMin() throws EmptyQExcptn;
  Comparable findMin() throws EmptyQExcptn;
  void makeEmpty();
  boolean isEmpty();
}
```

Heaps

A *heap* is a binary tree that stores `Comparable` objects such that each non-leaf node is less than both of its children. This means that the smallest object on the heap must occur at the root of the binary tree.



We actually *store* the heap as a linear array:

1	5	6	18	20	7	12	50	30	22	25	8	10	13

Notice that if the bottom level of the binary tree is not complete then it is filled from the left.

Parents and Children

Suppose the objects are stored in an array `items`. For reasons outlined later, we will not use `items[0]` to store an object, but start the storage at `items[1]`.

Then the following things are easy to see:

- The root of the tree is `items[1]`
- The left-hand child of `items[i]` is `items[2*i]`
- The right-hand child of `items[i]` is `items[2*i+1]`
- The parent of `items[i]` is `items[i/2]`.

Insertion

How can we insert an element into a heap?

The obvious answer is that it should be added to the end of the array, thus creating a binary tree with one more element. Unfortunately, the resulting binary tree is usually no longer a heap, and so we must restore the heap property.

This is done by an operation known as `percolateUp(int pos)` whereby the element in position `pos` is moved up the binary tree until the heap property is restored.

```
private void percolateUp(int pos) {

  Comparable tmp = items[pos];

  while (pos > 1 && tmp.lessThan(items[pos/2]) {
    items[pos] = items[pos/2];
    pos /= 2;
  }
  items[pos] = tmp;

}
```

The operation of `deleteMin()` returns and removes the root of the binary tree works.

If the root is removed from the tree, then there will be a "hole" left where the root was. Initially this hole is filled by replacing the element with the last element in the array (so that the array has no holes in it).

Once again, this will usually destroy the heap property of the binary tree, and it must be repaired by "percolating down" the offending element.

```
private void percolateDown(int pos) {

  Comparable tmp = items[pos];
  int smallest = pos;

  while (2*pos <= numItems) {
    smallest = 2*pos;
    if (2*pos+1 <= numItems)  {
      if (items[2*pos+1].lessThan(items[2*pos]))
         smallest = 2*pos+1;
    }
    if (items[smallest].lessThan(tmp)) {
      items[pos] = items[smallest];
      pos = smallest;
    }
    else break;
  }

  items[pos] = tmp;
}
```

Prim's algorithm

It is now easy to see how to implement Prim's algorithm. The objects that we will store in the priority queue will be *pairs* of the form $(v, k(v))$. The values $k(v)$ will always contain the length of the shortest edge connecting v to the spanning tree that is being grown. This value will be the value used by the priority queue for its comparisons.

The spanning tree being constructed will again be represented by an array π where, as before, $\pi(v)$ contains the parent of v in the spanning tree.

To initialize the algorithm, we set $k(s) = 0$, and add $(s, k(s))$ to the priority queue.

The repetitive step

At every stage of the algorithm, the pair $(u, k(u))$ with the lowest value of $k(u)$ is extracted from the priority queue. If the vertex u is already in the spanning tree, then it has already been fully examined. Otherwise, u gets incorporated into the spanning tree with the current value of $\pi(u)$ as its parent. Then we examine every edge (u, v) leading from u:

1. If v is already in the spanning tree, then we have already considered it, and nothing further needs to be done.

2. If v is not in the spanning tree, then we have to determine if the edge (u, v) should cause an update in the priority and parent of v. If the weight $w(u, v)$ of the edge (u, v) is less than $k(v)$ then we should set $k(v) = w(u, v)$ and $\pi(v) = u$.

This can be done in two ways—either another array can be used to keep track of where v lies in the priority queue, and the value updated, or we simply insert the pair $(v, w(u, v))$ into the priority queue.

Complexity of Prim

The complexity of Prim's algorithm is dominated by the heap operations.

Every vertex is extracted from the priority queue at some stage, hence the `deleteMin()` operations in the worst case take time $O(V \lg V)$.

Also, every edge is examined at some stage in the algorithm and each edge examination potentially causes an `insert()` operation. Hence in the worst case these operations take time $O(E \lg V)$.

Therefore the total time is

$$O(V \lg V + E \lg V) = O(E \lg V)$$

Priority-first search

Let us generalize the ideas behind this implementation of Prim's algorithm.

Consider the following very general graph-searching algorithm. We will later show that by choosing different specifications of the priority we can make this algorithm do very different things. This algorithm will produce a *priority-first search tree*.

The key-values or priorities associated with each vertex are stored in an array called k.

Then we select a source vertex s for the search, put $k(s) = 0$, and add $(s, k(s))$ to the priority queue.

PFS repetitive step

At every stage of the algorithm, the pair $(u, k(u))$ with the lowest value of $k(u)$ is extracted from the priority queue. If the vertex u is already in the spanning tree, then it has already been fully examined.

Otherwise, u gets incorporated into the spanning tree with the current value of $\pi(u)$ as its parent. Then we examine every edge (u, v) leading from u. For each such edge there are two possibilities:

1. If v is already in the spanning tree, then we have already considered it, and nothing further needs to be done.

2. If v is not in the spanning tree, then we have to determine if the edge (u, v) should cause an update in the priority and parent of v. If the value PRIORITY is less than $k(v)$ then we set $k(v) = PRIORITY$ and $\pi(v) = u$.

By changing the precise definition of PRIORITY we get a whole variety of searches.

Prim's algorithms is PFS

Prim's algorithm can be expressed as a priority-first search by observing that the priority of a vertex is the weight of the shortest edge joining the vertex to the rest of the tree.

This is achieved in the code above by simply replacing the string PRIORITY by

$$w(u, v)$$

At any stage of the algorithm:

• The vertices extracted from the priority queue form the spanning tree so far.

• For each vertex v not yet incorporated into the spanning tree, the value $k(v)$ gives the length of the shortest edge from v to a vertex in the spanning tree, and $\pi(v)$ holds the number of this vertex.

Shortest paths

Let G be a directed weighted graph; we have already defined the shortest path (that should be the lightest path); we denote the length of the shortest path from v to w by $\delta(v, w)$.

Let $s \in V(G)$ be a specified vertex called the *source* vertex.

The *single-source shortest paths* problem is to find the shortest path from s to every other vertex in the graph (as opposed to the *all-pairs shortest paths problem*, where we must find the distance between every pair of vertices).

Dijkstra's algorithm

Dijkstra's algorithm is a famous single-source shortest paths algorithm suitable for the cases when the weights are all non-negative.

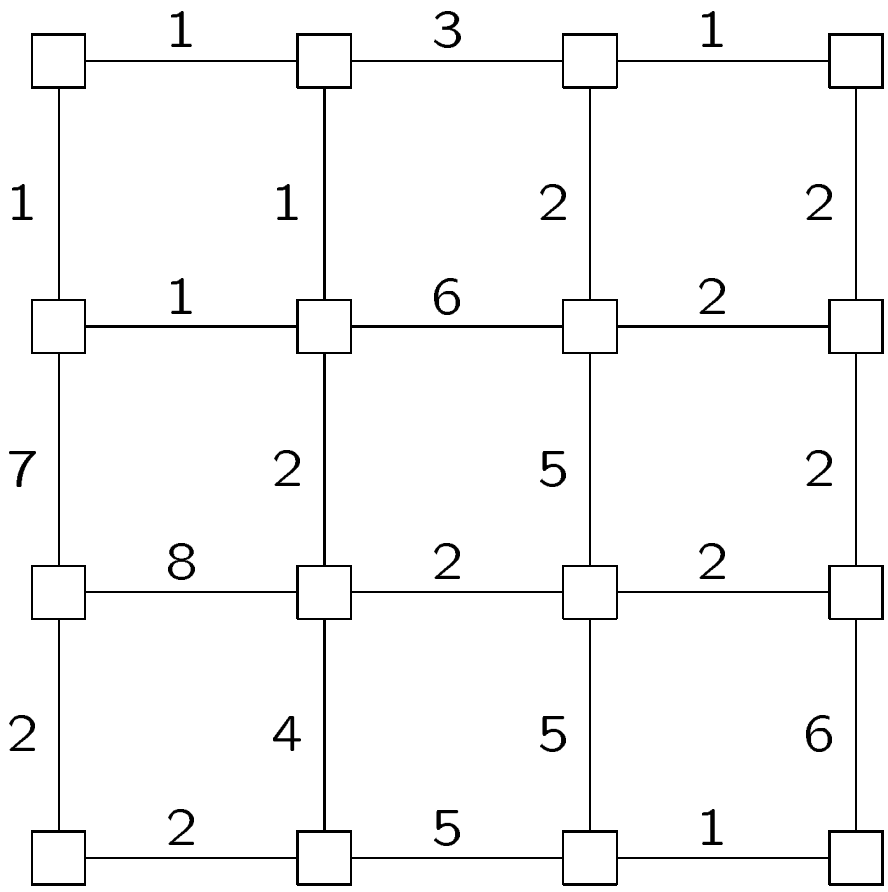
Dijkstra's algorithm can be implemented as a priority-first search by taking the priority of a vertex v to be the shortest path from s to v whose intermediate vertices lie in the priority-first search tree.

This can be implemented as a PFS by replacing PRIORITY with

$$k(u) + w(u, v)$$

At the end of the search, the array k contains the lengths of the shortest paths from the source vertex s.

Dijkstra's algorithm in action



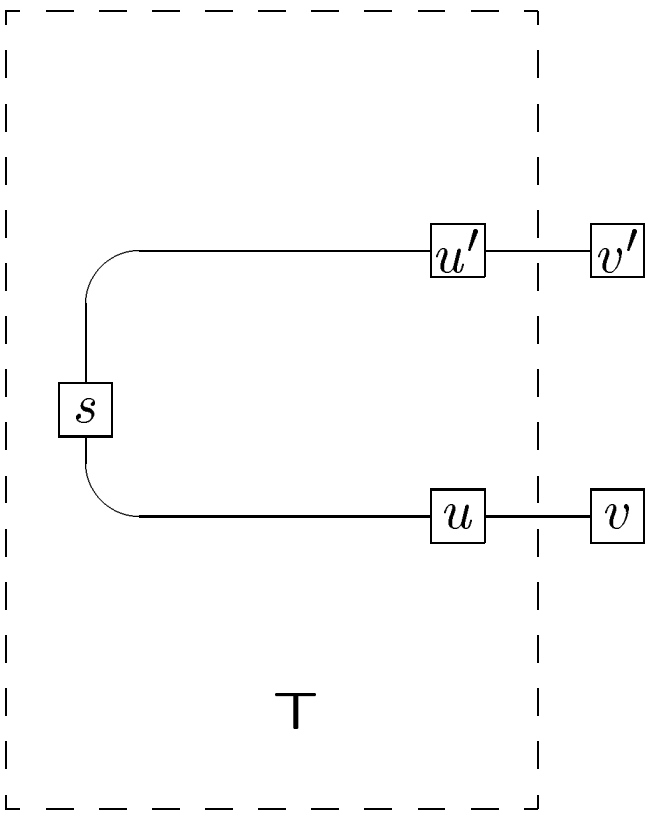
Proof of correctness

It is fairly easy to prove that Dijkstra's algorithm is correct by proving the following claim

- At the time that a vertex v is removed from the priority queue and placed into the priority first spanning tree T,

$$k(v) = \delta(s, v).$$

To prove this claim we consider the moment at which v is removed from the priority queue.



Proof

Suppose v has just been removed from the priority queue, and that $\pi(v) = u$. Then

$$k(v) = \delta(s, u) + w(u, v)$$

If this value is not the true value of $\delta(s, v)$ then there must be a shorter path from s to v. This shorter path must leave T at some stage, say the edge (u', v'). Then consider the value $k(v')$—because the edge (u', v') was examined when u' was added to T, the value of this key is at most $\delta(s, u') + w(u', v')$ This value however is less than the length of the shortest path from s to v and hence is less than the key value for v, which is impossible.

Therefore this situation cannot occur and we conclude that there cannot be any shorter paths from s to v and the result is true.

Complexity of PFS

The complexity of this search is easy to calculate — the main loop is executed V times, and each `deleteMin()` operation takes $O(\lg V)$ yielding a total time of $O(V \lg V)$ for the extraction operations.

During all V operations of the main loop we examine the adjacency list of each vertex exactly once — hence we make E calls, each of which may cause an `insert()` to be performed. Hence we do at most $O(E \lg V)$ work on these operations.

Therefore the total is

$$O(V \lg V + E \lg V) = O(E \lg V).$$

Relaxation

Consider the following property of Dijkstra's algorithm.

• At any stage of Dijkstra's algorithm the following inequality holds:

$$\delta(s, v) \leq k(v)$$

This is saying that the k array always holds a collection of *upper bounds* on the actual values that we are seeking. We can view these values as being our "best estimate" to the value so far, and Dijkstra's algorithm as a procedure for systematically improving our estimates to the correct values.

The fundamental step in Dijkstra's algorithm, where the bounds are altered is when we examine the edge (u, v) and do the following operation

$$k(v) = \min\{k(v), k(u) + w(u, v)\}$$

This is called *relaxing* the edge (u, v).

Relaxation schedules

Consider now an algorithm that is of the following general form:

- Initially an array d is initialized to have $d(s) = 0$ for some source vertex s and $d(v) = \infty$ for all other vertices
- A sequence of edge relaxations is performed, possibly altering the values in the array d.

We observe that the value $d(v)$ is always an upper bound for the value $\delta(s, v)$ because relaxing the edge (u, v) will either leave the upper bound unchanged or replace it by a better estimate from an upper bound on a path from $s \to u \to v$.

Dijkstra's algorithm is a particular schedule for performing the edge relaxations that guarantees that the upper bounds converge to the exact values.

Negative edge weights

Dijkstra's algorithm cannot be used when the graph has some negative edge-weights (why not? find an example).

In general, no algorithm for shortest paths can work if the graph contains a cycle of negative total weight (because a path could be made arbitrarily short by going round and round the cycle). Therefore the question of finding shortest paths makes no sense if there is a negative cycle.

However, what if there are some negative edge weights but no negative cycles?

The Bellman-Ford algorithm is a relaxation schedule that can be run on graphs with negative edge weights. It will either *fail* in which case the graph has a negative cycle and the problem is ill-posed, or will finish with the lengths of the shortest paths from s in the array d.

Bellman-Ford algorithm

The initialization step is as described above. Let us suppose that the weights on the edges are given by the function w.

Then consider the following relaxation schedule:

for $i = 1$ **to** $|V(G)| - 1$ **do**
    for each edge $(u, v) \in E(G)$ **do**
        RELAX(u,v)
    end for each
end for

Finally we make a single check to determine if we have a failure:

for each edge $(u, v) \in E(G)$ **do**
    if $d(v) > d(u) + w(u, v)$ **then** FAIL
    end if
end for each

Complexity of Bellman-Ford

The complexity is particularly easy to calculate in this case because we know exactly how many relaxations are done — namely $E(V-1)$, and adding that to the final failure check loop, and the initialization loop we see that Bellman-Ford is $O(EV)$

There remains just one question — how does it work?

To answer this, let us consider some of the properties of relaxation in a graph with no negative cycles.

Property 1 Consider an edge (u, v) that lies on the shortest path from s to v. If the sequence of relaxations includes relaxing (u, v) at a stage when $d(u) = \delta(s, u)$, then $d(v)$ is set to $\delta(s, v)$ and never changes after that.

Correctness of Bellman-Ford

Once convinced that Property 1 holds it is now simple to see that the algorithm is correct for graphs with no negative cycles.

Consider any vertex v and let us examine the shortest path from s to v, namely

$$s \sim v_1 \sim v_2 \cdots v_k \sim v$$

Now at the initialization stage $d(s) = 0$ and it always remains the same. After one pass through the main loop the edge (s, v_1) is relaxed and by Property 1, $d(v_1) = \delta(s, v_1)$ and it remains at that value. After the second pass the edge (v_1, v_2) is relaxed and after this relaxation we have $d(v_2) = \delta(s, v_2)$ and it remains at this value.

As the number of edges in the path is at most $|V(G)| - 1$, after all the loops have been performed $d(v) = \delta(s, v)$.

All-pairs shortest paths

Now we turn our attention to constructing a complete table of shortest distances, which must contain the shortest distance between any pair of vertices.

If the graph has no negative edge weights then we could simply make V runs of Dijkstra's algorithm, at a total cost of $O(VE \lg V)$, whereas if there are negative edge weights then we could make V runs of the Bellman-Ford algorithm at a total cost of $O(V^2E)$.

The two algorithms we shall examine both use the adjacency matrix representation of the graph, hence are most suitable for dense graphs. Recall that for a weighted graph the weighted adjacency matrix A has $w(i, j)$ as its ij-entry, where $w(i, j) = \infty$ if i and j are not adjacent.

A dynamic programming method

Dynamic programming is a general algorithmic technique for solving problems that can be characterised by two features:

- The problem is broken down into a collection of smaller subproblems
- The solution is built up from the stored values of the solutions to all of the subproblems

For the all-pairs shortest paths problem we define the simpler problem to be

"What is the length of the shortest path from vertex i to j that uses at most m edges?"

We shall solve this for $m = 1$, then use that solution to solve for $m = 2$, and so on ...

The initial step

We shall let $d_{ij}^{(m)}$ denote the distance from vertex i to vertex j along a path that uses at most m edges, and define $D^{(m)}$ to be the matrix whose ij-entry is the value $d_{ij}^{(m)}$.

As a shortest path between any two vertices can contain at most $V - 1$ edges, the matrix $D^{(V-1)}$ contains the table of all-pairs shortest paths.

Our overall plan therefore is to use $D^{(1)}$ to compute $D^{(2)}$, then use $D^{(2)}$ to compute $D^{(3)}$ and so on.

The case $m = 1$

Now the matrix $D^{(1)}$ is easy to compute — the length of a shortest path using at most one edge from i to j is simply the weight of the edge from i to j. Therefore $D^{(1)}$ is just the adjacency matrix A.

The inductive step

What is the smallest weight of the path from vertex i to vertex j that uses at most m edges? Now a path using at most m edges either be

1. A path using less than m edges

2. A path using exactly m edges, composed of a path using $m-1$ edges from i to an auxiliary vertex k and the edge (k, j).

We shall take the entry $d_{ij}^{(m)}$ to be the lowest weight path from the above choices.

Therefore we get

$$d_{ij}^{(m)} = \min\left(d_{ij}^{(m-1)}, \min_{0 \le k < V}\{d_{ik}^{(m-1)} + w(k,j)\}\right)$$

$$= \min_{0 \le k < V}\{d_{ik}^{(m-1)} + w(k,j)\}$$

Example

Consider the weighted graph with the following weighted adjacency matrix:

$$A = D^{(1)} = \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & \infty & \infty \\ 10 & \infty & 0 & \infty & \infty \\ \infty & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix}$$

Let us see how to compute an entry in $D^{(2)}$, suppose we are interested in the $(0, 2)$ entry:

Then we see that

$0 \to 0 \to 3$ has cost $0 + 11 = 11$
$0 \to 1 \to 3$ has cost $\infty + 4 = \infty$
$0 \to 2 \to 3$ has cost $11 + 0 = 11$
$0 \to 3 \to 3$ has cost $2 + 6 = 8$
$0 \to 4 \to 3$ has cost $6 + 6 = 12$

The minimum of all of these is 8, hence the $(0, 2)$ entry of $D^{(2)}$ is set to 8.

Computing $D^{(2)}$

$$\begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & \infty & \infty \\ 10 & \infty & 0 & \infty & \infty \\ \infty & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix} \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & \infty & \infty \\ 10 & \infty & 0 & \infty & \infty \\ \infty & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 4 & 8 & 2 & 5 \\ 1 & 0 & 4 & 3 & 7 \\ 10 & \infty & 0 & 12 & 16 \\ 3 & 2 & 6 & 0 & 3 \\ 16 & \infty & 6 & \infty & 0 \end{pmatrix}$$

If we multiply two matrices $AB = C$, then we compute

$$c_{ij} = \sum_{k=0}^{k=V-1} a_{ik} b_{kj}$$

If we replace the multiplication $a_{ik} b_{kj}$ by addition $a_{ik} + b_{kj}$ and replace summation Σ by the minimum min then we get

$$c_{ij} = \min_{k=0}^{k=V-1} a_{ik} + b_{kj}$$

which is precisely the operation we are performing to calculate our matrices.

The remaining matrices

Proceeding to compute $D^{(3)}$ from $D^{(2)}$ and A, and then $D^{(4)}$ from $D^{(3)}$ and A we get:

$$D^{(3)} = \begin{pmatrix} 0 & 4 & 8 & 2 & 5 \\ 1 & 0 & 4 & 3 & \boxed{6} \\ 10 & \boxed{14} & 0 & 12 & \boxed{15} \\ 3 & 2 & 6 & 0 & 3 \\ 16 & \infty & 6 & \boxed{18} & 0 \end{pmatrix}$$

$$D^{(4)} = \begin{pmatrix} 0 & 4 & 8 & 2 & 5 \\ 1 & 0 & 4 & 3 & 6 \\ 10 & 14 & 0 & 12 & 15 \\ 3 & 2 & 6 & 0 & 3 \\ 16 & \boxed{20} & 6 & 18 & 0 \end{pmatrix}$$

A new matrix "product"

Recall the method for computing $d_{ij}^{(m)}$, the (i,j) entry of the matrix $D^{(m)}$ using the method similar to matrix multiplication.

$d_{ij}^{(m)} \leftarrow \infty$
for $k = 0$ **to** $V - 1$ **do**
    $d_{ij}^{(m)} = \min(d_{ij}^{(m)}, d_{ik}^{(m-1)} + w(k,j))$
end for

We will use $\star$ to denote this new matrix product.

Then we have

$$D^{(m)} = D^{(m-1)} \star A$$

Hence it is an easy matter to see that we can compute as follows:

$$D^{(2)} = A \star A \qquad D^{(3)} = D^{(2)} \star A \ldots$$

Complexity of this method

The time taken for this method is easily seen to be $\Theta(V^4)$ as it performs V matrix "multiplications" each of which involves a triply nested **for** loop with each variable running from 1 to V.

However we can reduce the complexity of the algorithm by remembering that we do not need to compute *all* the intermediate products $D^{(1)}$, $D^{(2)}$ and so on, but we are only interested in $D^{(V-1)}$. Therefore we can simply compute:

$$D^{(2)} = A \star A$$

$$D^{(4)} = D^{(2)} \star D^{(2)}$$

$$D^{(8)} = D^{(4)} \star D^{(4)}$$

Therefore we only need to do this operation at most $\lg V$ times before we reach the matrix we want.

Floyd-Warshall

The Floyd-Warshall algorithm uses a different dynamic programming formalism.

For this algorithm we shall define $d_{ij}^{(k)}$ to be the length of the shortest path from i to j whose intermediate vertices all lie in the set $\{0, \ldots, k\}$.

As before, we shall define $D^{(k)}$ to be the matrix whose (i, j) entry is $d_{ij}^{(k)}$.

The initial case

What is the matrix $D^{(-1)}$ — the entry $d_{ij}^{(-1)}$ is the length of the shortest path from i to j with *no* intermediate vertices. Therefore $D^{(-1)}$ is simply the adjacency matrix A.

The inductive step

For the inductive step we assume that we have constructed already the matrix $D^{(k-1)}$ and wish to use it to construct the matrix $D^{(k)}$.

Let us consider all the paths from i to j whose intermediate vertices lie in $\{0, 1, \ldots, k\}$. There are two possibilities for such paths

(1) The path does not use vertex k
(2) The path does use vertex k

The shortest possible length of all the paths in category (1) is given by $d_{ij}^{(k-1)}$ which we already know.

If the path does use vertex k then it must go from vertex i to k and then proceed on to j, and the length of the shortest path in this category is $d_{ik}^{(k-1)} + d_{kj}^{(k-1)}$.

The overall algorithm

The overall algorithm is then simply a matter of running V times through a loop, with each entry being assigned as the minimum of two possibilities. Therefore the overall complexity of the algorithm is just $O(V^3)$.

$D^{(-1)} \leftarrow A$
for $k = 0$ **to** $V - 1$ **do**
    for $i = 0$ **to** $V - 1$ **do**
        for $j = 0$ **to** $V - 1$ **do**
            $d_{ij}^{(k)} = \min(d_{ij}^{(k-1)}, d_{ik}^{(k-1)} + d_{kj}^{(k-1)})$
        end for j
    end for i
end for k

At the end of the procedure we have the matrix $D^{(V-1)}$ whose (i, j) entry contains the length of the shortest path from i to j, all of whose vertices lie in $\{0, 2, \ldots, V - 1\}$—in other words, the shortest path.

Example

Consider the weighted directed graph with the following adjacency matrix:

$$D^{(-1)} = \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & \infty & \infty \\ 10 & \infty & 0 & \infty & \infty \\ \infty & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix}$$

Let us see how to compute $D^{(0)}$

$$D^{(0)} = \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & & \\ 10 & \infty & 0 & & \\ \infty & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix}$$

To find the $(1,3)$ entry of this matrix we have to consider the paths through the vertex 0 — is there a path from $1-0-3$ that has a better value than the current path? If so, then that entry is updated.

The entire sequence of matrices

$$D^{(1)} = \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & \boxed{3} & \boxed{7} \\ 10 & \infty & 0 & \boxed{12} & \boxed{16} \\ \boxed{3} & 2 & 6 & 0 & 3 \\ \infty & \infty & 6 & \infty & 0 \end{pmatrix}$$

$$D^{(2)} = \begin{pmatrix} 0 & \infty & 11 & 2 & 6 \\ 1 & 0 & 4 & 3 & 7 \\ 10 & \infty & 0 & 12 & 16 \\ 3 & 2 & 6 & 0 & 3 \\ \boxed{16} & \infty & 6 & \boxed{18} & 0 \end{pmatrix}$$

$$D^{(3)} = \begin{pmatrix} 0 & \boxed{4} & \boxed{8} & 2 & \boxed{5} \\ 1 & 0 & 4 & 3 & \boxed{6} \\ 10 & \boxed{14} & 0 & 12 & \boxed{15} \\ 3 & 2 & 6 & 0 & 3 \\ 16 & \boxed{20} & 6 & 18 & 0 \end{pmatrix}$$

$$D^{(4)} = \begin{pmatrix} 0 & 4 & 8 & 2 & 5 \\ 1 & 0 & 4 & 3 & 6 \\ 10 & 14 & 0 & 12 & 15 \\ 3 & 2 & 6 & 0 & 3 \\ 16 & 20 & 6 & 18 & 0 \end{pmatrix}$$

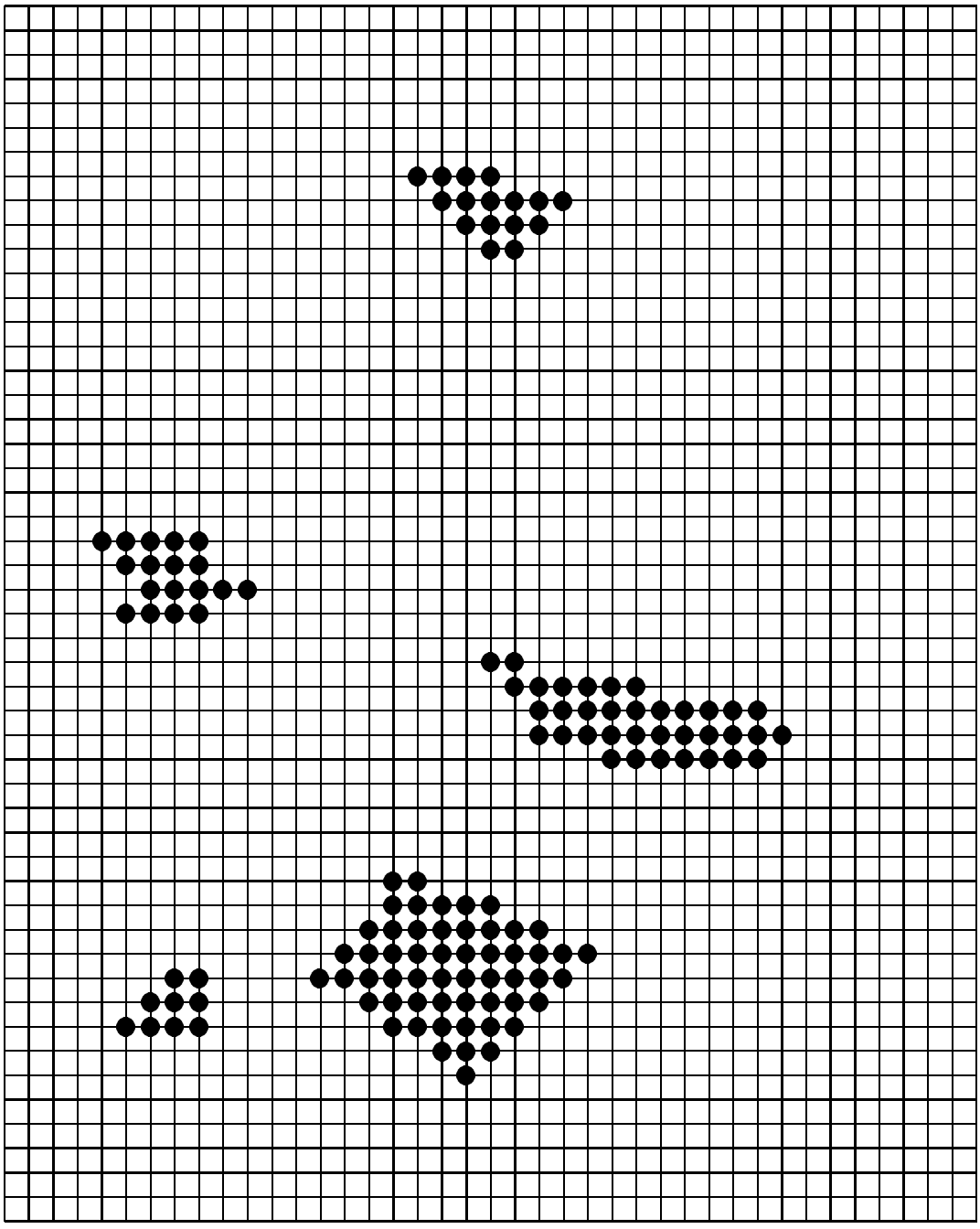
Finding the actual shortest paths

In both of these algorithms we have not addressed the question of actually finding the paths themselves.

For the Floyd-Warshall algorithm this is achieved by constructing a further sequence of arrays $P^{(k)}$ whose (i, j) entry contains a predecessor of j on the path from i to j. As the entries are updated the predecessors will change — if the matrix entry is not changed then the predecessor does not change, but if the entry does change, because the path originally from i to j becomes re-routed through the vertex k, then the predecessor of j becomes the predecessor of j on the path from k to j.

The A^* algorithm

Suppose we have a robot which moves in a 2-dimensional space. Our task is to find the shortest route through a cluttered landscape from a fixed starting point to a goal position.



A graph problem

This can easily be turned into a graph problem — the vertices of the graph are the points of the grid not covered by obstacles and we join two vertices if the robot can move between them in a single horizontal or vertical move.

In some sense, this graph is a "state-space graph for a robot" and it is clear that we need to find a shortest path from the source vertex s to the goal vertex g in this graph.

If we regard the 2-dimensional plane on which the robot is moving as infinite in size, then the graph that we (mentally) construct is also infinite, although we will obviously only be explicitly considering a small part of it.

What about Dijkstra?

Dijkstra's algorithm does solve the single-source shortest paths problem, but only in graphs sufficiently small to be completely examined.

The algorithm finds shortest paths from the source vertex s to *all* other vertices, making no use of the fact that we are only interested in the one vertex g.

In a state-space graph that may be extremely large (for example a 1000×1000 grid) a vast amount of unnecessary work is done. Of course the algorithm can be modified to stop as soon as g is reached, and we will definitely have found the shortest path.

This will work and in the absence of any further knowledge is the best that can be done.

Using estimates

The search can be dramatically improved however if we use some additional knowledge to concentrate the search in the more promising areas of the graph.

One of the ways in which the search can be improved is if we have some sort of *estimate* of the distance from each vertex to the goal. Then whenever we encounter a vertex v, we have an estimate for the length of the shortest path to the goal that passes through v — the sum of the known length already travelled to v and the estimated distance from v to the goal.

It seems natural then to examine the most promising looking paths first — in other words to give these paths a higher priority.

A priority first search

We formalize the above idea:

Let G be an edge-weighted graph, and let s, g be two vertices of G. Furthermore suppose that we have an estimating function e such that for any vertex v, $e(v, g)$ is an estimate of the distance from v to g.

We conduct a version of priority-first search; we will maintain arrays d and π such that $d(v)$ is the length of the shortest path that we have found from s to v, and $\pi(v)$ is the immediate predecessor of v on that path.

To initialize things we set $d(s) = 0$ and $\pi(s)$ to be undefined. Then we insert the pair $(s, e(s, g))$ into the priority queue.

The priority-first search

Then the search proceeds as follows:

- $x =$ `pq.deleteMin()`
- If $x = g$ then stop
- For each neighbour y of x compute the value

$$d(x) + w(x, y)$$

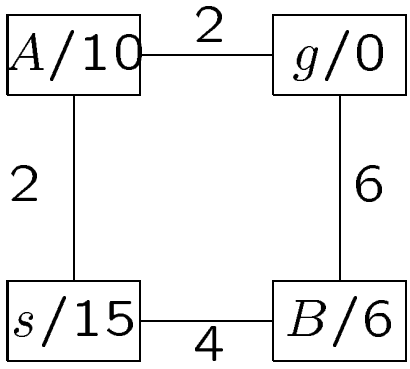
- If this value is less than $d(y)$ then set $d(y)$ to $d(x) + w(x, y)$, set $\pi(y)$ to x and insert y into the priority queue with priority

$$d(x) + w(x, y) + e(y, g).$$

Observe that unlike Dijkstra's algorithm, it is conceivable that a vertex can re-enter the priority queue even if it has been dequeued before.

Finding optimal solutions

It is very easy to see that this algorithm does not always find optimal solutions:



In this example immediately after the first node s is examined, the priority queue will contain

$$(B, 10) \quad (A, 12)$$

so B will be next examined, after which the priority queue will contain

$$(g, 10) \quad (A, 12)$$

and then g will be at the head of the queue, so the algorithm will stop and declare that $s - B - g$ is the optimal route.

Underestimates

The problem here is that the estimate of the remaining distance from A is such a bad overestimate that the paths starting $s - A -?$ are never examined. This problem can be completely eliminated by using an *underestimate* for the estimated remaining distance.

Theorem If $e(v, g) \leq \delta(v, g)$ then the procedure above is guaranteed to produce an optimal path.

Proof At the moment that the algorithm terminates the vertex g has the lowest priority. Consider a true shortest path from s to g. At every stage of the algorithm there is some vertex v of this path on the priority queue for which $d(v) = \delta(s, v)$. The priority of this vertex is less than or equal to $\delta(s, g)$ and therefore the priority of g is actually equal to $\delta(s, g)$ and the shortest path has been found.

More on underestimates

The quality of the search depends highly on the quality of the underestimates. We can consider what happens at the two extremes.

If the underestimates are exactly correct, then the algorithm never takes a wrong step — it simply moves directly along a true shortest path from s to g.

If the underestimates are as bad as they can be, (that is, every estimate for the remaining distance is zero) then the algorithm simply reduces to Dijkstra's algorithm.

One way to view this algorithm is that the underestimates turn Dijkstra's algorithm into an *informed search*.

The A^* algorithm for motion planning

This priority-first search using underestimates is called the A^* algorithm.

It is primarily used in the Artificial Intelligence area, because AI problems tend to involve some state-space searching with specific goals.

For the motion planning problem described above, a natural underestimate of the distance from any point to the goal is given by the Euclidean distance (that is, the length of a straight line from the point to the goal, ignoring obstacles and the fact that the robot is constrained to move horizontally and vertically).

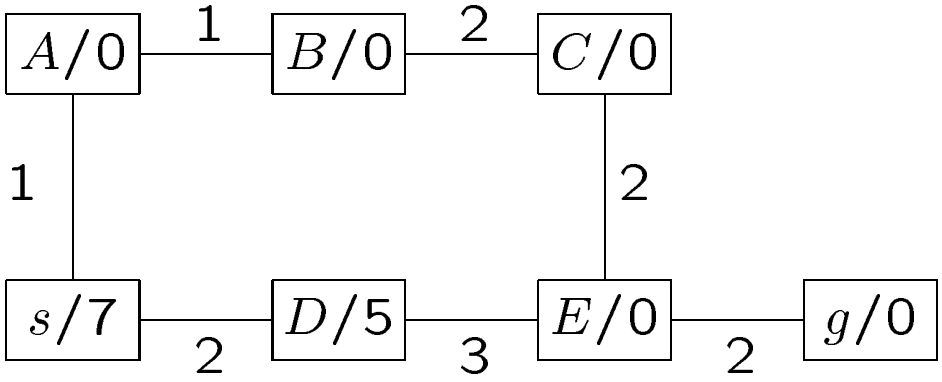
Properties of A^*

Dijkstra's algorithm has the nice property that whenever a vertex v is removed from the priority queue we are guaranteed that

$$d(v) = \delta(s, v).$$

This means that once a vertex is dequeued we never need to put it back into the queue because a shorter path to it has been found.

Unfortunately this is not true for the A^* algorithm as the following example shows:



The search proceeds s, A, B, C and then E is dequeued when $d(E) = 6$, whereas $\delta(s, E) = 5$.

Monotone functions

The reason that this occurs is because the underestimate function is badly chosen. In particular consider the estimates for the vertices D and E: with $w(D, E) = 3$ the values

$$e(D, g) = 5 \qquad e(E, g) = 0$$

seem foolish—the underestimate $e(E, g)$ is clearly far too low.

The underestimating function e is said to be *monotone* if for all vertices x and y we have

$$e(x, g) \leq w(x, y) + e(y, g)$$

If the underestimating function e is monotone, then

$$d(v) = \delta(s, v)$$

when v is dequeued.

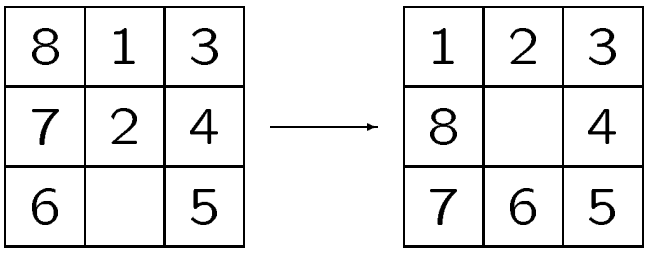
The Euclidean distance underestimate function is monotone.

An AI example

Consider the following puzzle known as the 8-puzzle (or sometimes the 9-puzzle). In its physical form it consists of 8 plastic tiles labelled 1, 2, ..., 8 arranged in a 3×3 frame. The tiles can slide horizontally or vertically into the empty space. The puzzle is shown in its finished or *goal* state.

1	2	3
8		4
7	6	5

The aim of the puzzle is to start with a "random" initial configuration and end at the goal configuration.



The state-space graph

The state-space graph of the 8-puzzle has $9! = 2 \times 181440$ states—however only 181440 of these are in the same connected component as the goal configuration.

The distance of these states from the goal is given by the following table.

Dist	Number	Dist	Number	Dist	Number
1	4	2	8	3	8
4	16	5	32	6	60
7	72	8	136	9	200
10	376	11	512	12	964
13	1296	14	2368	15	3084
16	5482	17	6736	18	11132
19	12208	20	18612	21	18444
22	24968	23	19632	24	22289
25	13600	26	11842	27	4340
28	2398	29	472	30	148

An A^* solution

The large number of intermediate states shows that a BFS would have to have a very large queue, and examine a large number of states. Therefore this puzzle is an ideal application for the A^* algorithm.

There are two obvious underestimating functions that can be used:

- $e(S, G) =$ the number of tiles out of position
- $e(S, G) =$ the sum of the Manhattan distances of each tile from its goal position

Both of them will work, but the second is far more efficient and leads to quite quick solutions of the A^* puzzle. A Java applet that solves the puzzle is on the ALG site.