NB Procedure ① and procedure ② yielded DIFFERENT spanning trees of (V, E) as we had lots of choices as to which edges to delete or add respectively. We thus see that a spanning tree of a connected graph is not unique unless of course, the original graph is itself a tree.

Kruskal's algorithm

Task If each edge of a connected graph (V, E) comes w/ a particular cost, describe an algorithm that finds the spanning tree of (V, E) w/ minimal cost.

Def Let (V, E) be an undirected graph. A cost function $c: E \to \mathbb{R}$ on the set E of edges of the graph is a function that assigns to each edge e of the graph a real number $c(e)$.

Let $c: E \to \mathbb{R}$ be a cost function on the set E of edges of a graph (V, E). Given any subset $S \subset E$, we define the cost on S to be $c(S) = \sum_{e \in S} c(e)$, the sum of the costs of all elements of S.

<u>Def</u> Let (V,E) be a connected graph w/ cost function $c: E \to \mathbb{R}$. A spanning tree (V, E_M) is said to be <u>minimal</u> (with respect to the cost function) if $\forall$ (V, E_T) a spanning tree of (V,E)
$c(E_M) \leq c(E_T)$.

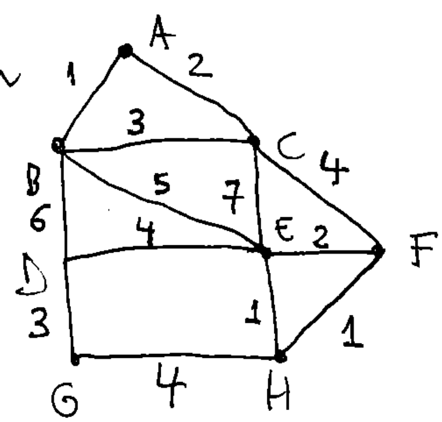
<u>Kruskal's Algorithm for finding minimal spanning trees:</u>

Let (V,E) be a connected graph w/ an associated cost function $c: E \to \mathbb{R}$.

1. Start w/ $(V, \emptyset)$, the subgraph of (V,E) consisting of all the vertices of (V,E) and no edges.
2. List all edges in E in a <u>queue</u> so that the cost of the edges is non-decreasing in the queue, i.e. if $e, e' \in E$ and if $c(e) < c(e')$, then e precedes e' in the queue.
3. Take edges successively from the front of the queue, and determine whether or not the addition of that edge to the current subgraph will create a cycle (circuit). If a circuit is created by this addition, discard the edge; otherwise, add it to the subgraph. Continue until the queue is emptied.

We will first do an example, and after the example we will prove Kruskal's algorithm yields a spanning tree that is indeed minimal.

<u>Example</u>: Consider



The cost function here is

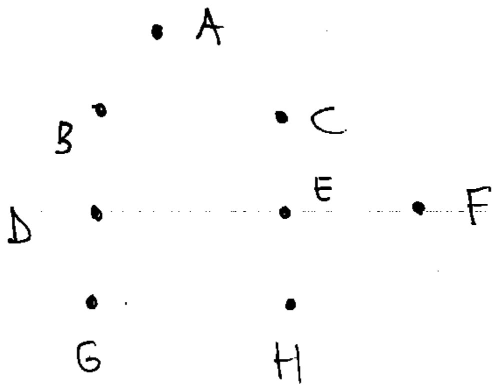
$c(AB) = 1$ $\quad c(CE) = 7$
$c(AC) = 2$ $\quad c(CF) = 4$
$c(BC) = 3$ $\quad c(EF) = 2$
$c(BD) = 6$ $\quad c(EH) = 1$
$c(BE) = 5$ $\quad c(FH) = 1$
$c(DE) = 4$ $\quad c(DG) = 3$

$c(GH)=4$

We can also use a table format to write down the cost function $c: E \longrightarrow \mathbb{R}$

AB	AC	BC	BD	BE	DE	CE	CF	EF	EH	FH	DG	GH
1	2	3	6	5	4	7	4	2	1	1	3	4

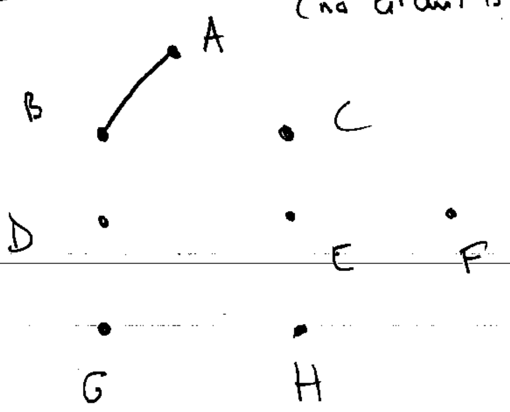
We start w/ $(V, \emptyset)$. This is step 0 of the algorithm. (The starting state).



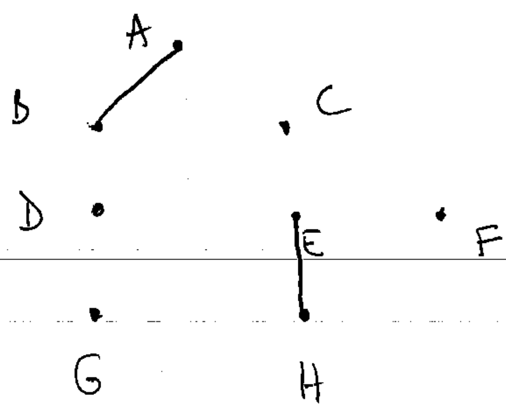
We list the edges in a queue so that the cost is non-decreasing

AB, EH, FH, AC, EF, BC, DG, DE, CF, GH, BE, BD, CE

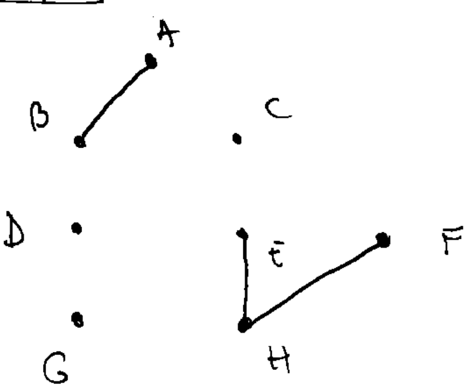
Step 1 We can add AB (no circuit is formed)



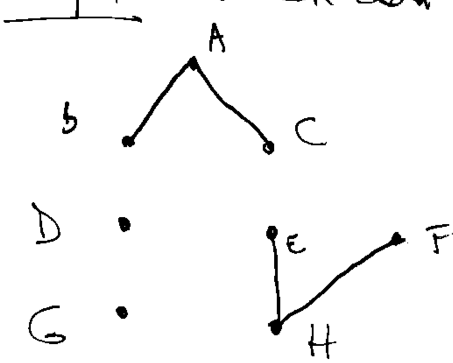
Step 2 We can add EH



Step 3 We can add FH



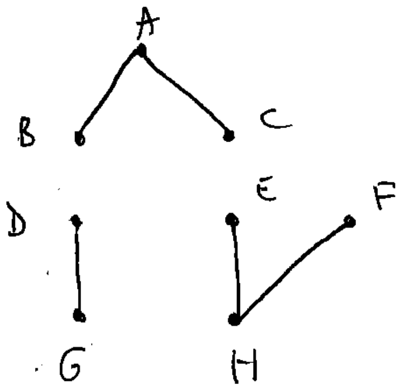
Step 4 We can add AC



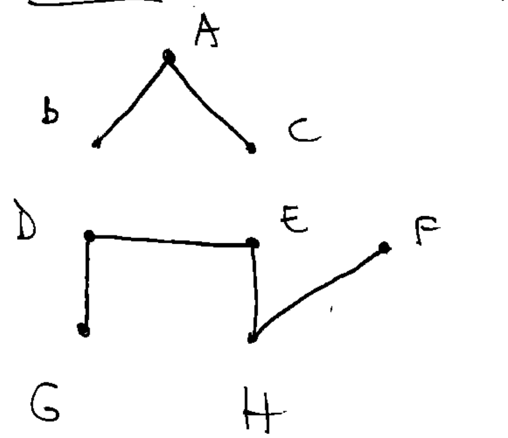
Step 5 We cannot add edge EF because we would create circuit EFHE, so EF gets discarded.

Step 6 We cannot add edge BC because we would create circuit ABCA, so BC gets discarded.

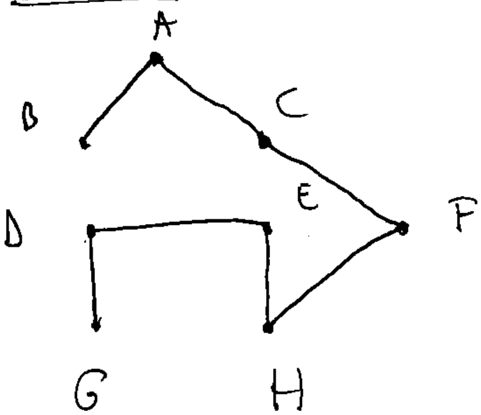
Step 7 We can add DG.



Step 8 We can add DE.



Step 9 We can add CF.



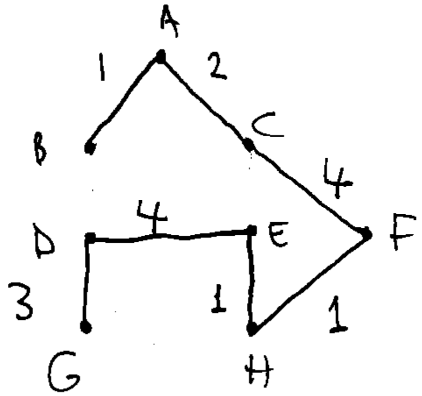
Step 10 We cannot add GH because we would create circuit DEHGD.

Step 11 We cannot add BE because we would create circuit BEHFCAB.

Step 12 We cannot add BD because we would create circuit BDEHFCAB.

Step 13 We cannot add edge CE because we would create circuit CEHFC.

The minimal spanning tree given by Kruskal's algorithm is thus:



min cost $c(E_T) = 1+2+4+1+1+4+3 = 16$

Now that we have some intuition about the Kruskal algorithm, let us prove that it always yield a spanning tree that is indeed minimal.

<u>Proposition</u> Let (V,E) be a connected graph with associated cost function $c: E \to \mathbb{R}$. Kruskal's algorithm yields a spanning tree of (V,E).

<u>Proof</u> Since an edge is added from the queue only if no circuit is formed, we conclude the subgraph (V,E') of (V,E) produced by the Kruskal algorithm must be acyclical (i.e. contain no circuits.). To prove (V,E') is a spanning tree of (V,E), we must show (V,E') is connected. Assume not, then (V,E') has components $(V_1,E_1'), (V_2,E_2'), \ldots, (V_m,E_m')$ for $m \geq 2$. (V,E) is connected, however $\Rightarrow \exists$ edge $e_{ij} \in E$ for $1 \leq i,j \leq m$, $i \neq j$ s.t. adding edge e_{ij} connects (V_i,E_i') and (V_j,E_j'), but edge e_{ij} <u>could not</u> have possibly created a circuit when considered in the queue $\Rightarrow\!\Leftarrow$ (V,E') cannot have more than one connected component $\Rightarrow (V,E')$ is connected. (q.e.d.)

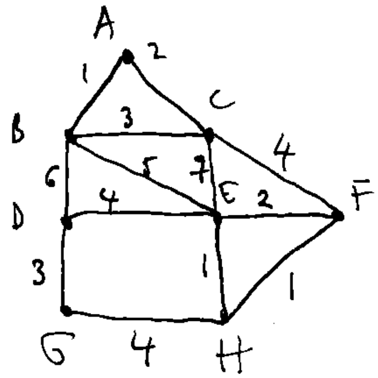
<u>Proposition</u> Let (V, E) be a connected graph w/ associated cost function $c: E \to \mathbb{R}$. Kruskal's algorithm yields a minimal spanning tree of (V, E).

<u>Proof</u> We already showed in the previous proposition that Kruskal's algorithm yields a spanning tree. Now we have to show that spanning tree is minimal w.r.t. $c: E \to \mathbb{R}$. Let (V, E') be the spanning tree given by the algorithm. If $(V, E') = (V, E)$, i.e. if the original connected graph is a tree, then there is nothing to prove. Assume $(V, E') \neq (V, E)$ i.e. (V, E) contains some circuit. Let all the edges of (V, E) be $e_1, e_2, \ldots, e_m$ that we label s.t. $c(e_i) \leq c(e_j)$ $\forall$ $1 \leq i < j \leq m$. In other words, $c(e_1) \leq c(e_2) \leq \ldots \leq c(e_{m-1}) \leq c(e_m)$. Kruskal's algorithm chooses the lowest cost $\#(V) - 1$ edges from $e_1, e_2, \ldots, e_m$ such that the resulting subgraph is a spanning tree of (V, E). Therefore, if (V, E'') is any other spanning tree of (V, E), then $c(E') \leq c(E'')$. (q.e.d.)

<u>Def</u>: Let (V, E) be a connected graph w/ associated cost function $c: E \to \mathbb{R}$. Let (V, E') be the minimal spanning tree of (V, E) produced by Kruskal's algorithm. (V, E') is called the <u>Kruskal tree</u>.

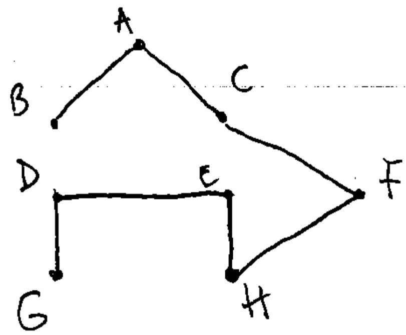
<u>NB</u> If two or more edges have the same cost, then we can reshuffle them in the queue used to determine the Kruskal tree. Therefore, the Kruskal tree might not be unique. In the example we used to illustrate Kruskal's algorithm

We see this scenario at work:

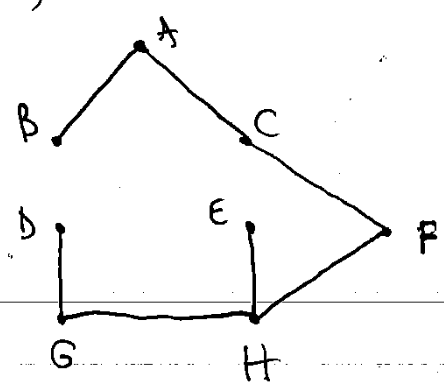


$C(DE) = C(GH) = 4.$

We used the queue AB, EH, FH, AC, EF, BC, DG, DE, CF, GH, BE, BD, CE to produce the Kruskal tree



whereas the queue AB, EH, FH, AC, EF, BC, DG, GH, CF, DE, BE, BD, CE would have produced the Kruskal tree



which has the same cost.

Remarks ① Joseph Kruskal published this algorithm that bears his name in 1956, two years after he finished his PhD at Princeton. Kruskal is known for work in computer science, combinatorics, and statistics.

② The cost of an edge is sometimes called the weight of that edge.

(3) Kruskal's algorithm starts w/ a disconnected graph $(V, \emptyset)$ and adds edges until the graph becomes connected and a tree, thus a spanning tree. In other words, until the last addition of an edge, the graph is disconnected.