Objectives

- Minimum Spanning Tree
- Union-Find Data Structure
- Clustering

Review

- What does the acronym MST stand for?
  - What is an MST?
- What are some algorithms to find the MST?
- What did we prove about the intersection of cycles and cut sets?
- How do we prove the following:
  - **Cut property.** Let *S* be any subset of nodes, and let *e* be the min cost edge with exactly one endpoint in *S*. Then the MST T* contains *e*.
  - Pf. (exchange argument)
    - Suppose there is an MST T* that does not contain *e*
      - What do we know about T, by defn?
      - What do we know about the nodes *e* connects?

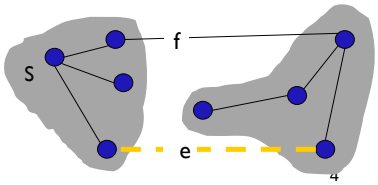
Proving Cut Property: OK to Include Edge

- Simplifying assumption: All edge costs c_e are distinct.
- **Cut property.** Let *S* be any subset of nodes, and let *e* be the min cost edge with exactly one endpoint in *S*.
  Then the MST T* contains *e*.
- Pf. (exchange argument)
  - Suppose there is an MST T* that does not contain *e*
    - What do we know about T, by defn?
    - What do we know about the nodes *e* connects?

Proving Cut Property: OK to Include Edge

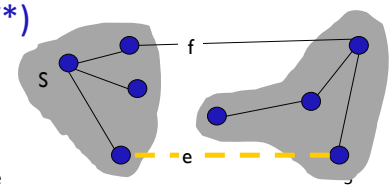
- Cut property. Let *S* be any subset of nodes, and let *e* be the min cost edge with exactly one endpoint in *S*. Then the MST T* contains *e*.
- Pf. (exchange argument)
  - Suppose there is an MST T* that does not contain *e*
  - Adding *e* to T* creates a cycle *C* in T*
  - Edge *e* is in cycle *C* and in cutset corresponding to *S*
    - ⇒ there exists another edge, say *f*, that is in both *C* and S's cutset

Which means?



Proving Cut Property: OK to Include Edge

- Cut property. Let *S* be any subset of nodes, and let *e* be the min cost edge with exactly one endpoint in *S*. Then the MST T* contains *e*.
- Pf. (exchange argument)
  - Suppose there is an MST T* that does not contain *e*
  - Adding *e* to T* creates a cycle *C* in T*
  - Edge *e* is in cycle *C* and in cutset corresponding to *S*
    - ⇒ there exists another edge, say *f*, that is in both *C* and S's cutset
  - T' = T* ∪ { e } - { f } is also a spanning tree
  - Since $c_e < c_f$, cost(T') < cost(T*)
  - This is a contradiction. ■



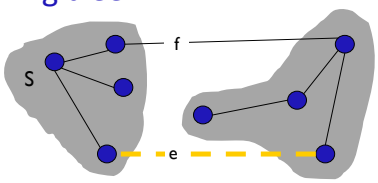
Proving Cycle Property: OK to Remove Edge

- Simplifying assumption: All edge costs c_e are distinct
- **Cycle property.** Let *C* be any cycle in *G*, and let *f* be the max cost edge belonging to *C*. Then the MST T* does not contain *f*.

Ideas about approach?

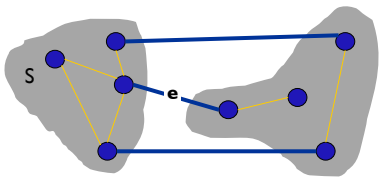
Cycle Property: OK to Remove Edge

- Cycle property. Let *C* be any cycle in *G*, and let *f* be the max cost edge belonging to *C*. Then the MST T* does not contain *f*.
- Pf. (exchange argument)
  - Suppose *f* belongs to T*
  - Deleting *f* from T* creates a cut *S* in T*
  - Edge *f* is both in the cycle *C* and in the cutset *S*
    - $\Rightarrow$ there exists another edge, say *e*, that is in both *C* and S
  - T' = T* $\cup$ { *e* } - { *f* } is also a spanning tree
  - Since $c_e < c_f$, cost(T') < cost(T*)
  - This is a contradiction. ▪

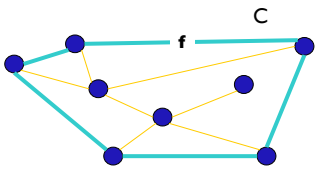


Summary of What We Proved

- Simplifying assumption: All edge costs c_e are distinct
  - → MST is unique
- Cut property. Let *S* be any subset of nodes, and let *e* be the min cost edge with exactly one endpoint in *S*. Then MST contains *e*.
- Cycle property. Let *C* be any cycle, and let *f* be the max cost edge belonging to *C*. Then MST does not contain *f*.



Cut Property: e is in MST



Cycle Property: f is **not** in MST

Prim's Algorithm

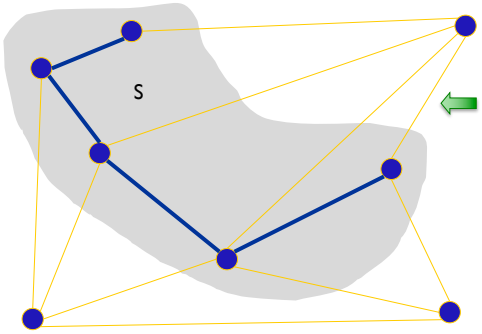
[Jarník 1930, Dijkstra 1957, Prim 1959]

- Start with some root node *s* and greedily grow a tree T from *s* outward.
- At each step, add the cheapest edge *e* to T that has exactly one endpoint in T.

How can we prove its correctness?

Prim's Algorithm: Proof of Correctness

- Initialize S to be any node
- Apply ***cut property*** to S
  - Add min cost edge (v, u) in *cutset* corresponding to S, and add one new explored node u to S



Ideas about implementation?

Implementation: Prim's Algorithm

Similar to Dijkstra's algorithm

- Maintain set of explored nodes *S*
- For each unexplored node *v*, maintain attachment cost a[v] → cost of cheapest edge *v* to a node in *S*

Running Time?

```
foreach (v ∈ V) a[v] = ∞
Initialize an empty priority queue Q
foreach (v ∈ V) insert v onto Q
Initialize set of explored nodes S = ϕ
while (Q is not empty)
   u = delete min element from Q
   S = S ∪ { u }
   foreach (edge e = (u, v) incident to u)
       if ((v ∉ S) and (ce < a[v]))
           decrease priority a[v] to ce
```


Implementation: Prim's Algorithm

Similar to Dijkstra's algorithm

- Maintain set of explored nodes *S*
- For each unexplored node *v*, maintain attachment cost a[v] → cost of cheapest edge *v* to a node in *S*

O(m log n) with a heap

```
foreach (v ∈ V) a[v] = ∞                    O(n)
Initialize an empty priority queue Q
foreach (v ∈ V) insert v onto Q              O(n logn)
Initialize set of explored nodes S = ϕ
while (Q is not empty)                       O(n)
   u = delete min element from Q             O(log n)
   S = S ∪ { u }
   foreach (edge e = (u, v) incident to u)   O(deg(u))
       if ((v ∉ S) and (ce < a[v]))
           decrease priority a[v] to ce      O(log n)
```

Kruskal's Algorithm [1956]

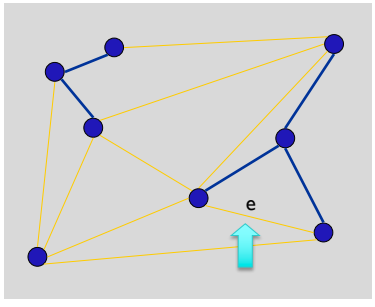
- Start with T = ϕ
- Consider edges in *ascending order of cost*
- Insert edge *e* in T unless doing so would create a cycle
  - Add edge as long as "compatible"

How can we prove algorithm's correctness?

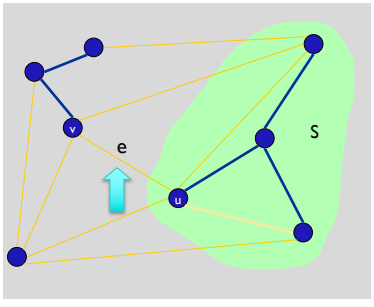
Kruskal's Algorithm: Proof of Correctness

What is tricky about implementing Kruskal's algorithm?

- Consider edges in ascending order of weight
- Case 1: If adding *e* to T creates a cycle, discard *e* according to ***cycle property*** (*e* must be max weight)
- Case 2: Otherwise, insert *e* = (u, v) into T according to ***cut property*** where *S* = set of nodes in u's *connected component*



Case 1



Case 2

Implementing Kruskal's Algorithm

What is tricky about implementing Kruskal's algorithm?

How do we know when adding an edge will create a cycle?
- What are the properties of a graph/its nodes when adding an edge will create a cycle?

UNION-FIND DATA STRUCTURE

Union-Find Data Structure

- Keeps track of a graph as edges are added
  - Cannot handle when edges are deleted
- Maintains disjoint sets
  - E.g., graph's connected components
- Operations/API:
  - `Find(u)`: returns name of set containing u
    - How utilized to see if two nodes are in the same set?
    - Goal implementation: **O(log n)**
  - `Union(A, B)`: merge sets A and B into one set
    - Goal implementation: **O(log n)**

Best darn Union-Find Data Structure

Implementing Kruskal's Algorithm

- Using the **union-find** data structure
  - Build set T of edges in the MST
  - Maintain set for each connected component

Costs?

```
Sort edge weights so that c1 ≤ c2 ≤ ... ≤ cm
T = {}
foreach (u ∈ V) make a set containing singleton u

for i = 1 to m                    are u and v in different connected components?
   (u,v) = ei
   if (u and v are in different sets)
      T = T ∪ {ei}
      merge the sets containing u and v
return T                          merge two components
```

Looking Ahead

- Wiki: 4.5-4.7
- PS7 – next Friday