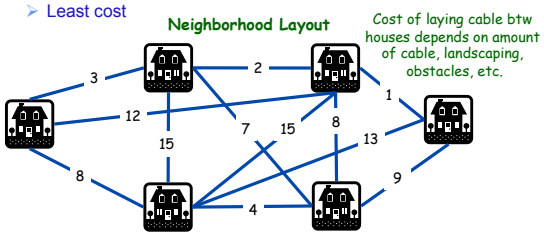
Objectives

- Minimum Spanning Tree
- Union-Find data structure
- Clustering

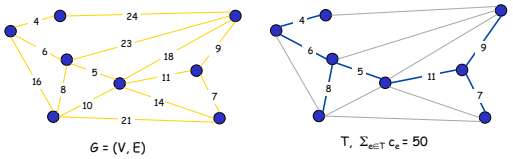
Review: Laying Cable

- Comcast knows how to make money and how to save money
- They want to lay cable in a neighborhood
  - Reach all houses
  - Least cost



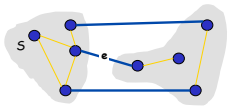
Review: Minimum Spanning Tree

- Given a connected graph G = (V, E) with positive edge weights c_e, an MST is a subset of the edges $T \subseteq E$ such that T is a *spanning tree* whose sum of edge weights is *minimized*
  - Spanning tree: spans all nodes in graph

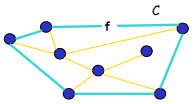


Review: Greedy Algorithms

- 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

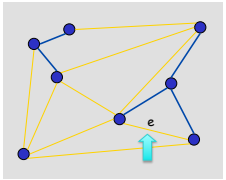
Kruskal's Algorithm [1956]

- 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"

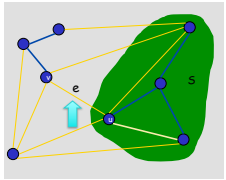
How can we prove algorithm's correctness?

Kruskal's Algorithm: Proof of Correctness

- 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:
  - **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 U-F 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 edges 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
```


Implementing Kruskal's Algorithm

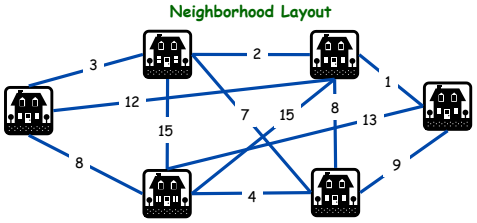
- Using best implementation of union-find
  - Sorting: O(m log n) ← $m \le n^2 \Rightarrow \log m$ is $O(\log n)$
  - Union-find: O(m α (m, n))
  - O(m log n) — essentially a constant

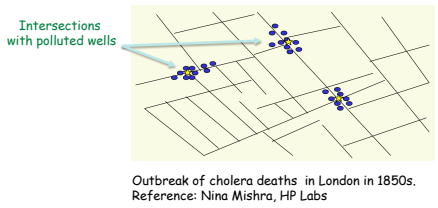
```
Sort edges weights so that c1 ≤ c2 ≤ ... ≤ cm
T = {}
foreach (u ∈ V) make a set containing singleton u

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
```


Limitations to Applying MST?

- Motivating Example: Comcast laying cable





Outbreak of cholera deaths in London in 1850s.
Reference: Nina Mishra, HP Labs

CLUSTERING

Clustering

- Given a set *U* of *n* objects (or points) labeled $p_1, \ldots, p_n$, classify into coherent groups
  - Example objects: photos, documents, micro-organisms
- Distance function. Numeric value specifying "closeness" of two objects

Clustering Problem

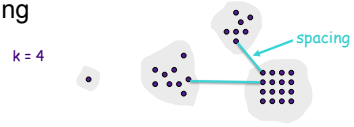
- Divide objects into clusters so that points in different clusters are far apart
- Applications
  - Routing in mobile ad hoc networks
  - Identify patterns in gene expression
  - Identifying patterns in web application use cases
    - Sets of URLs
  - Similarity searching in medical image databases
  - Skycat: cluster 109 sky objects into stars, quasars, galaxies

Clustering

- k-clustering. Divide objects into *k* non-empty groups
- Distance function. Assume it satisfies several natural properties
  - $d(p_i, p_j) = 0$ iff $p_i = p_j$ (identity of indiscernibles)
  - $d(p_i, p_j) \geq 0$ (nonnegativity)
  - $d(p_i, p_j) = d(p_j, p_i)$ (symmetry)

Clustering of Maximum Spacing

- k-clustering. Divide objects into *k* non-empty groups
- Spacing. Min distance between any pair of points in different clusters
- Clustering of maximum spacing. Given an integer *k*, find a *k*-clustering of maximum spacing



Ideas about Solving?

- Greedy algorithm?
- How relates to the minimum spanning tree?

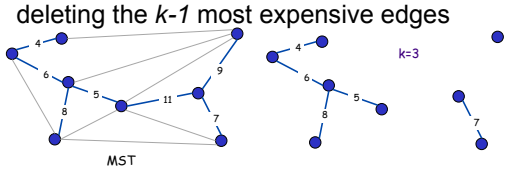
Greedy Clustering Algorithm

- Single-link *k*-clustering algorithm
  - Form a graph on the vertex set *U*, corresponding to *n* clusters
  - Find the closest pair of objects such that *each object is in a different cluster*, and add an edge between them
  - Repeat *n-k* times until there are exactly *k* clusters

How is this related to the MST?

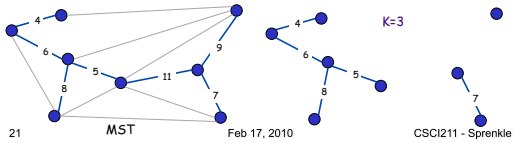
Greedy Clustering Algorithm

- Key observation. Same as Kruskal's algorithm
  - Except we stop when there are *k* connected components
- Remark. Equivalent to finding an MST and deleting the *k-1* most expensive edges



Greedy Clustering Algorithm: Analysis

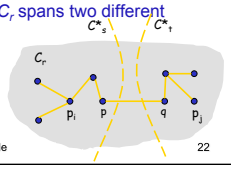
- Theorem. Let C denote the clustering C_1, …, C_k formed by deleting the *k-1* most expensive edges of a MST. C is a *k*-clustering of *max spacing*.
- Pf Intuition:
  - What can we say about C's spacing?
    - Within clusters and between clusters
  - What if C isn't optimal?
    - What does that mean about C's clusters vs (optimal) C*'s clusters?



Greedy Clustering Algorithm: Analysis

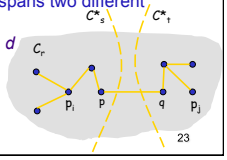
- Theorem. Let C denote the clustering C_1, …, C_k formed by deleting the *k-1* most expensive edges of a MST. C is a *k*-clustering of *maximum spacing*.
- Pf Sketch. Let C* denote some other clustering C^*_1, …, C^*_k. C* and C must be different; otherwise we're done.
  - The spacing of *C* is length *d* of (k-1)st most expensive edge
  - Let p_i, p_j be in the same cluster in *C* (say C_r) but different clusters in C*, say C^*_s and C^*_t
  - Some edge (*p*, *q*) on p_i-p_j path in C_r spans two different clusters in *C**

What do we know about *(p, q)*?



Greedy Clustering Algorithm: Analysis

- Theorem. Let C denote the clustering C_1, …, C_k formed by deleting the *k-1* most expensive edges of a MST. C is a *k*-clustering of *maximum spacing*.
- Pf. Let C* denote some other clustering C^*_1, …, C^*_k. C* and C must be different; otherwise we're done.
  - The spacing of *C* is length *d* of (k-1)st most expensive edge
  - Let p_i, p_j be in the same cluster in *C* (say C_r) but different clusters in C*, say C^*_s and C^*_t
  - Some edge (*p*, *q*) on p_i-p_j path in C_r spans two different clusters in *C**
  - All edges on p_i-p_j path have length ≤ *d* since Kruskal chose them
  - Spacing of C* is at most ≤ d since *p* and *q* are in different clusters

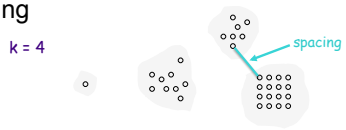


Assignments

- PS 4 due Friday
- Continue reading chapter 4

Problem: Clustering of Maximum Spacing

- k-clustering. Divide objects into *k* non-empty groups
- Spacing. Min distance between any pair of points in different clusters
- Clustering of maximum spacing. Given an integer *k*, find a *k*-clustering of maximum spacing



Our Proposed Solution

- Start with each node in its own cluster
- Sort edges by their distance, ascending
- For each edge, combine its nodes' clusters into one cluster until we have k clusters