AMAT 583, Lec 20, 11/7/19

<u>Today</u>: Single linkage clustering
- graphs
- dendrograms

<u>Review</u> (generalities about clustering)

The <u>input</u> X to a clustering method is a finite subset of $\mathbb{R}^n$, or more generally, a finite metric space.

The <u>output</u> is one of the following:

1) A (sub)partition of X
2) A hierarchical (sub)partion of X.

Methods that output a hierarchical (sub) partition are called <u>hierarchical clustering methods</u>.

<u>Recall</u>: A partition of X is a set P of non-empty subsets X such that each element of X is contained in exactly one element of P.

A subpartition P of X is a partition of a subset of X
$\Rightarrow$ each element of X is contained in at most one element of P.

A hierarchical (sub)partition of X is a collection $\{P_\alpha\}_{\alpha \in [0,\infty)}$ of (sub)partions of X such that if $\alpha \leq \beta$ and $A \in P_\alpha$, then $A \subset B$ for some $B \in P_\beta$. (I phrased this slightly differently, but equivalently, in the last lecture).

<u>Note</u>: Sometimes for simplicity, we'll consider hierarchical subpartitions indexed by $\mathbb{N} = \{0, 1, 2, \ldots\}$ rather than $[0,\infty)$.

Today, we'll focus on a simple and very natural clustering technique called <u>single</u> <u>linkage</u>.

Single linkage is closely connected to topology and has good mathematic properties. But it has some bad properties that make it useful only in special settings.

<u>Input</u>: Any finite metric space (X, d)

<u>Output</u>: A hierarchical partition of X.

To explain single linkage, we will need to define <u>graphs</u>.

Graphs

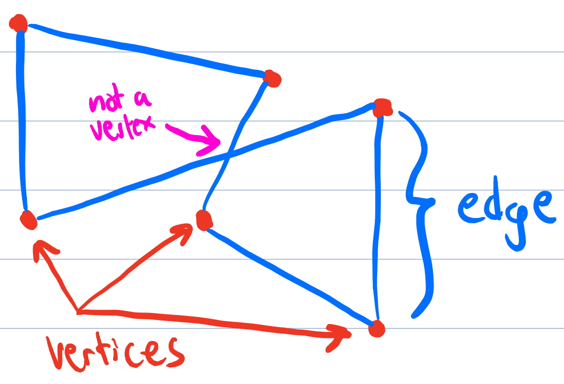
- very important constructions in computer science, mathematics, and statistics

- Note: These graphs are not the same as the graphs of of functions you've seen since high school.

We distinguish between directed and undirected graphs.

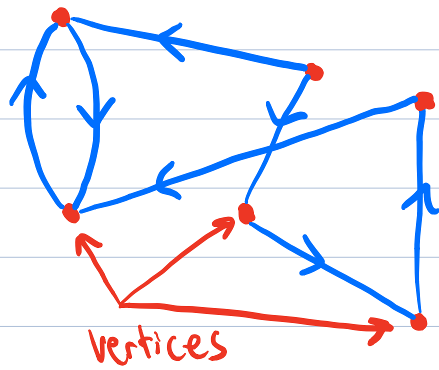
Intuitively, an undirected graph is a collection of points (vertices) with edges connecting them.

We can draw this in the plane like so



However, formally, we don't usually assume that the vertices live in $\mathbb{R}^2$.

A directed graph is a similar kind of object, but the edges are each assumed to have a direction from one vertex to another.



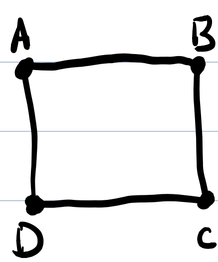
For now, we will be concerned only with undirected graphs, so we define only these formally.

Definition: An undirected graph is a pair (V, E) where

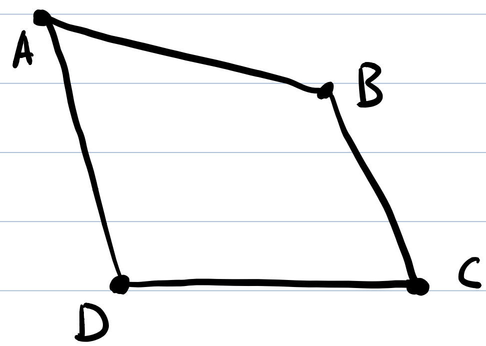
- V is any set (called the vertex set)
- E is a set of two element subsets of V (called the edge set). We will abuse notation slightly and write the two element set $\{A, B\}$ as $[A, B]$.

Example: $V = \{A, B, C, D\}$, $E = \{[A,B], [B,C], [C,D], [D,A]\}$.

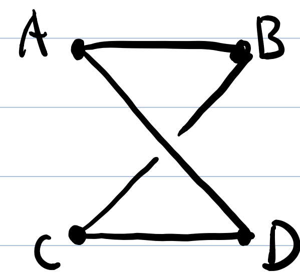
We can draw this as follows:



or like this:



or even like this:

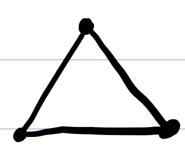


(there's no prescribed rule for where to place the vertices in the plane, though some choices are clearly more awkward than others.

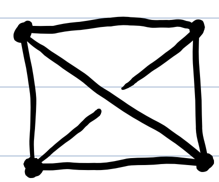
<u>Example</u>: A complete <u>graph</u> is one with an edge between every possible pair of vertices.

Complete graph on 2 vertices:

Complete graph on 3 vertices:

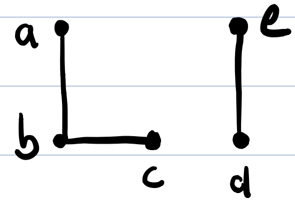


Complete graph on 4 vertices:



Example: A graph can have multiple "components."

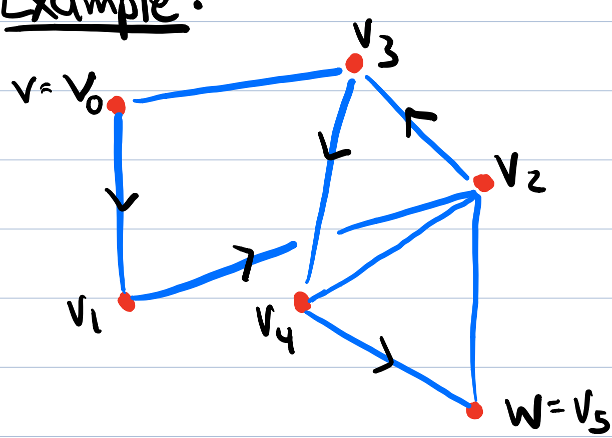
$V = \{a, b, c, d, e\}$.
$E = \{[a,b], [b,c], [d,e]\}$



Connected components of (undirected) graphs

For $G = (V, E)$ an undirected graph and $v, w \in V$, a **path** from v to w is a sequence of $n \geq 1$ vertices $v = v_1, v_2, \ldots, v_n = w$ such that for $1 \leq i \leq n-1$, $[v_i, v_{i+1}] \in E$.

Example:



Note: If $n=1$, then $v=v_1=v_n=v_w$ and the 1-element sequence v is a path from v to itself.

Define a relation $\sim$ on V by taking $v \sim w$ iff $\exists$ a path from v to w.

Proposition: $\sim$ is an equivalence relation.

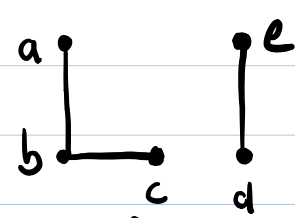
Proof is an easy exercise

A subgraph of a graph $G=(V,E)$ is a graph $G'=(V',E')$ with $V' \subset V$, $E' \subset E$.

Def: A connected component of G is a subgraph $G=(V',E')$ such that

1) V' is an equivalence class of $\sim$
2) $E' = \{(v,w) \in E \mid v,w \in V'\}$. ← That is, every edge between vertices in V' is included in E!

Example: Let $G=$



The connected components of G are

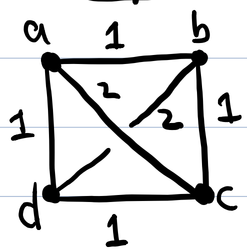
$G^1 = (\{a,b,c\}, \{[a,b],[b,c]\})$
$G^2 = (\{d,e\}, \{[e,d]\})$.

Let (X, d) be a finite metric space.

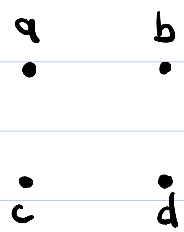
For simplicity, assume the metric is integer-valued.

For $z \in \mathbb{N} = \{0, 1, 2, 3, \ldots\}$, let $N_z(X)$ be the graph with:
- Vertex set X
- An edge $[x, y]$ included iff $d(x, y) \leq z$.

Example: $X = \{a, b, c, d\}$, with the metric given as follows

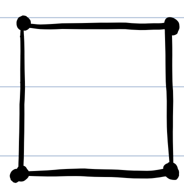


Then $N_0(X) =$

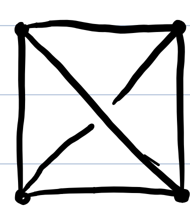


[no edges]

$N_1(X) =$



$N_2(X) =$



Note that if $y \leq z$, $N_y(X) \subset N_z(X)$.

We define the <u>single linkage clustering</u> of X $\{P_z\}_{z \in \mathbb{N}}$ by taking

$$P_z = \{X' \subset X \mid X' \text{ is the vertex set of a connected component of } N_z(X)\}$$