AMAT 583 Lec 22 11/14/19

Today: Dendrograms continued
Single linkage as a topological clustering method.
Average linkage.

Review

Def: A discrete hierarchical partition of X is a collection $P=\{P_\alpha\}_{\alpha\in\mathbb{N}}$ of partions of X such that if $\alpha\leq\beta$ and $A\in P_\alpha$, then $A\subset B$ for some $B\in P_\beta$.

"discrete" means $\alpha\in\mathbb{N}$.

Dendrograms

Let $P=\{P_z\}_{z\in\mathbb{N}}$ be a discrete hierarchical partition of a finite set X.

The (unlabeled) dendrogram of P consists of:

- A graph $D(P)=(V,E)$
- A function $L: V\to\mathbb{N}$

Specifically, $V=\{(S,z)\mid z\in\mathbb{N},\ S\in P_z\}$

so every element of every partition P_z corresponds to one vertex in the graph.

$$E = \{ [(S,z),(T,z+1)] \mid z \in \mathbb{N},\ S < T \}.$$

L is defined by $L(S,z) = z$.

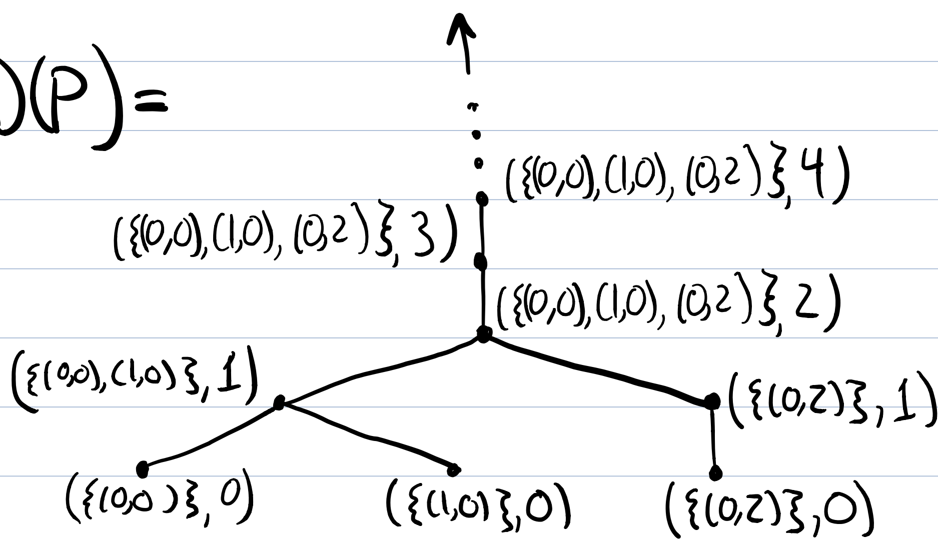
<u>Example (continuing from last time)</u>

$$X = \{(0,0),(1,0),(0,2)\}$$

$P = SL((X, d_1))$, i.e.

$$P_z = \begin{cases} \{\{(0,0)\},\{(1,0)\},\{(0,2)\}\} & \text{if } z=0 \\ \{\{(0,0),(1,0)\},\{(0,2)\}\} & \text{if } z=1 \\ \{\{(0,0),(1,0),(0,2)\}\} & \text{if } z \geq 2. \end{cases}$$

$D(P) =$



Remark: We can define the dendrogram of a discrete hierarchical sub-partition in the exactly same way.

Proposition: For any hierarchical subpartition P, $D(P)$ is a forest.

Proof: Exercise.

Trimming the dendrogram

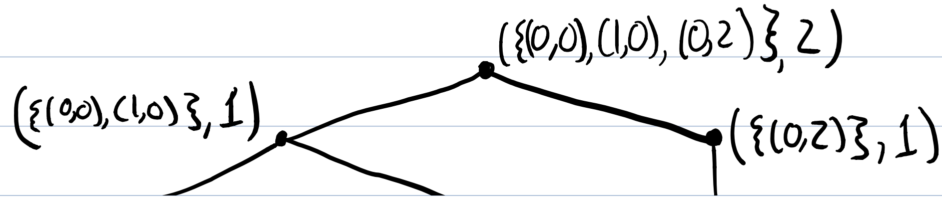
The dendrogram can be simplified in two ways, without loss of information

1) Since X is finite, there will be some smallest z_{top} such that $P_z = P_{z+1}$ for all $z \geq z_{top}$

   For instance, in the above example, $z_{top} = 2$.

   We usually only plot the subgraph of $D(P)$ consisting of vertices (S, z) with $z \leq z_{top}$, and all edges between such vertices.

   We'll denote this $\hat{D}(P)$

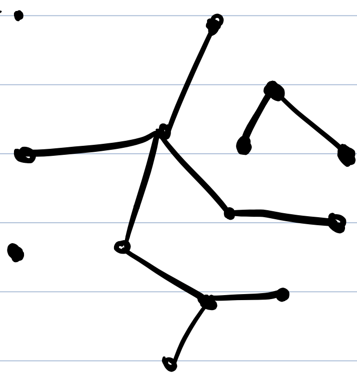


$(\{(0,0)\},0)$ $(\{(1,0)\},0)$ $(\{(0,2)\},0)$

2) For a graph $G=(V,E)$ and $v,w \in V$, we say v is indicent to w if $[v,w] \in E$.

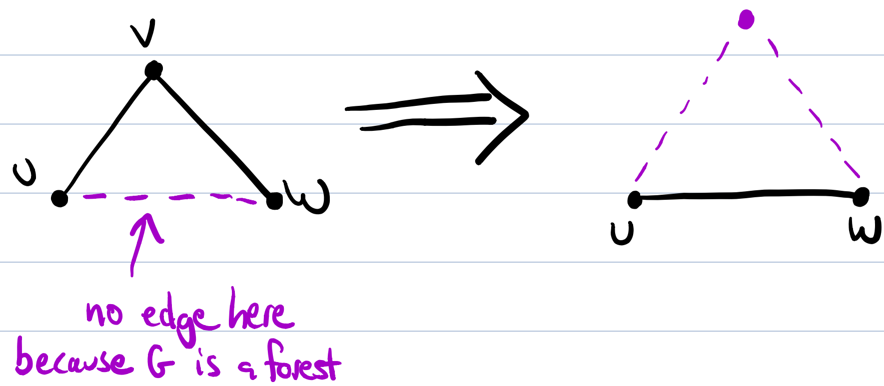
The degree of v, denoted $\deg(v)$, is the number of edges incident to w.

Exercise: Label each vertex of the following graph by its degree.



For $G=(V,E)$ any forest, there is a natural way of removing any vertices of degree 2 from G:

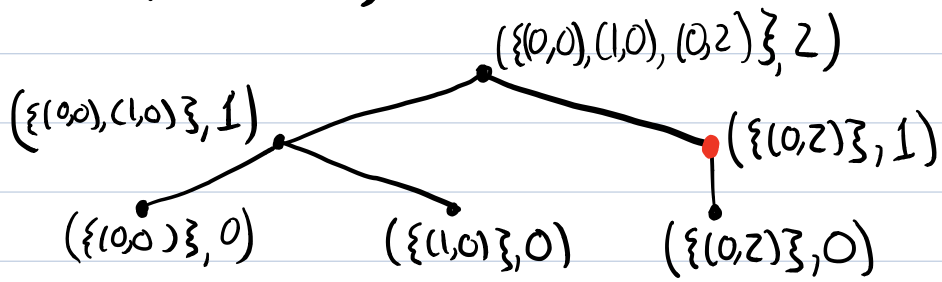
- Suppose v is a vertex of degree 2, incident to vertices u,w.
- We remove v, $[v,u]$, and $[v,w]$ from G.
- We add $[u,v]$ to G.



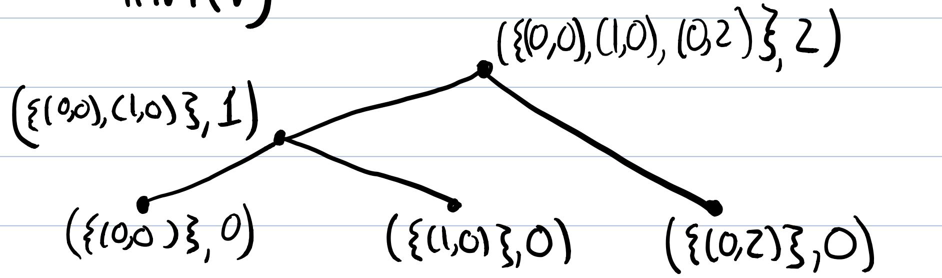
If v is a vertex in $\hat{D}(P)$ incident to $u, v,$ with $L(u) < L(v) < L(w)$, we remove V.

Let $\mathrm{Trim}(V)$ denote the graph obtained from $\hat{D}(P)$ in any order. Call V the trimmed dendrogram.

<u>Example</u>: $\hat{D}(P)$

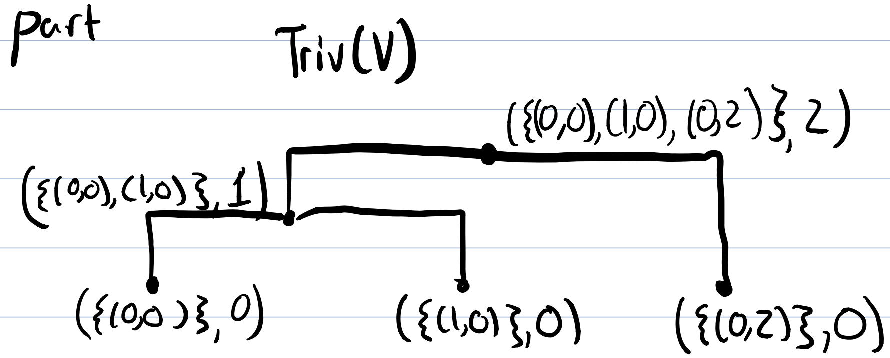


$\mathrm{Trim}(V)$

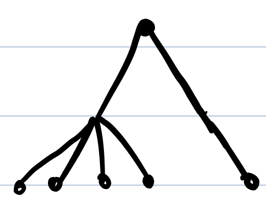


We visualize a hierarchical partition via its trimmed dendrogram.

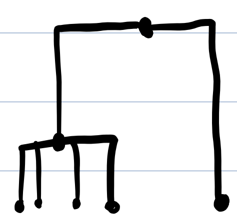
We usually put an "elbow" in each edge so that is drawn as a vertical part, followed by horizontal part



What if we have something like this?



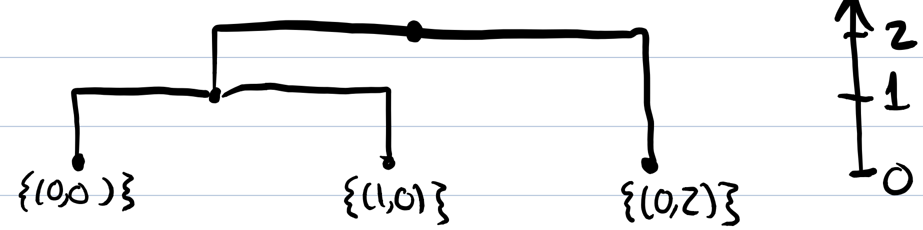
Then we draw:



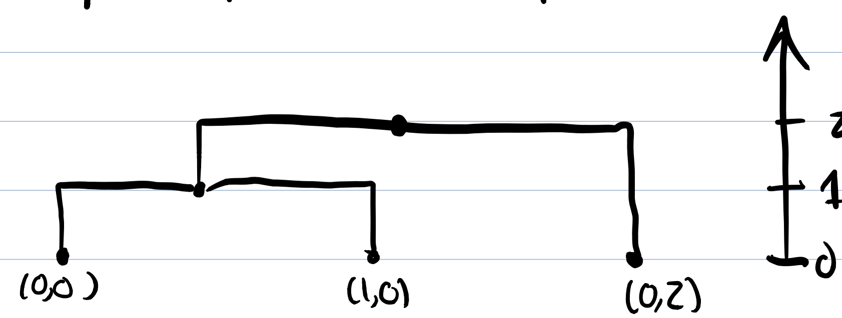
Labeling the dendrogram of a partition

For each vertex $(S,0)$, we label the vertex by S.

- We typically do not label any other vertices.
- We can include an indication of scale to the side



If the partition consists only of singletons at level 0, we can simplify the labels further. (This assumption is nearly always satisfied in practice.



This is the dendrogram that clustering software will produce

Dendrograms of non-discrete hierarchical partitions

Assume $P = \{P_\alpha\}_{\alpha \in [0,\infty)}$ is a partition of a finite set X.

Since X is finite, the partition can change only at a finite number of indices

$\alpha_1 < \alpha_2 < \alpha_3 < \cdots < \alpha_n$.

Restricting P to these indices gives the first n partitions in a discrete hierarchical partition.

Single linkage and Topology

Two (related) stories connecting single linkage to topology.

Geometric realization of a graph

- We draw graphs as geometric objects.
- This can be formalized via a construction called geometric realization.

Given a graph G construct the following topological space:

- We have one copy of