Lecture 2: Trees

Cuong Van Than

October 24, 2013

Trees. A graph (either directed or undirected) that has no cycle is called acyclic. An undirected tree is an undirected, connected acyclic graph. A directed graph is a (directed) tree if its underlying graph (i.e. the graph obtained when the direction of edges is ignored) is an undirected tree.

A rooted tree is a tree in which a vertex is distinguished as the root. Normally, when a tree is rooted, its edges are considered as directed, either away or toward the root.

Theorem 1. *The following properties hold for an (undirected) tree T:*

1. *Every two vertices are joined by a unique path.*
2. $|E(T)| = |V(T)| - 1$.
3. *If we add a new edge joining two non-adjacent vertices of T, then the resulting graph contains exactly one cycle.*

Proof. The first and third properties are straightforward. The second can be proved by induction. $\square$

A vertex of degree one is called a leaf; otherwise, it is an internal (or interior) vertex. An edge incident to a leaf is called a pendant leaf. An internal edge is an edge joining two internal vertices.

Theorem 2. *Every tree with at least two vertices have at least two leaves.*

Proof. The claim can be proved by either considering a maximal path (i.e. a path that is *not* part of another path with strictly larger length), or using the identity $\sum_{v\in V(T)} \deg(v) = 2|E(T)| = 2(|V(T)| - 1)$. $\square$

We say that an unrooted tree is binary if every internal vertex has degree three. A rooted tree is binary if every vertex has degree three, except for the root that has degree two. In other words, considering the edges of a rooted tree as being directed away from the root, and considering v a child of u if (u, v) is a (directed) edge, we say that a rooted tree is binary if every internal vertex has exactly two children.

Centers. Let G be a graph, and let u and v be two vertices of G. The distance from u to v, denoted as $d(u, v)$, is the length of a shortest path from u to v; if u and v are not connected, we set $d(u, v) = \infty$. The eccentricity $e(u)$ is defined as the maximum of $d(u, v)$ for all $v \in V(G)$. The radius $r(G)$ is the minium of eccentricity, and the diameter $d(G)$ is the maximum of eccentricity.

A point v is called a central point if $e(v) = r(G)$, and the center of G is defined as the set of central points.

Theorem 3. *The center of an undirected tree consists of either a single vertex or two adjacent vertices.*

Proof. We prove by induction on the number of vertices a tree T. If T has only one or two vertices, then the claim clearly holds. Otherwise, let T' be the tree obtained from T by deleting all the leaves of T. Observe that if $d_T(u, v) = e_T(u)$, then v is leaf of T. It follows that $e'_T(v) = e_T(v) - 1$ for every internal vertex of T, and hence a central point of T' is also a central point of T, and vice versa (note that a leaf of T cannot be central point, except for the case T has only two vertices). The claim now follows. $\square$

Cut-edges and cut-vertices. An edge of a graph G is called a cut-edge if its removal increases the number of connected components of G. Similarly, a vertex of G is called a cut-vertex if its removal increases the number of connected components of G.

Theorem 4. *A connected graph is a tree if and only if every edge is a cut-edge.*

X-trees. Let X be a nonempty set of labels. An (unrooted) X-tree is an ordered pair (T, ϕ), where T is an (unrooted) tree and ϕ is a map from X to $V(T)$ in which every vertex of degree at most two is labeled. Note that:

1. As ϕ is a map from X to V, a vertex of T can have more than one label.
2. By definition, every leaf or 2-degree vertex must have (at least) a label.
3. However, it is possible that a vertex with degree greater than two is labeled.

When no misunderstanding can arise, we refer to an X-tree (T, ϕ) as T for brevity.

A phylogenetic tree (on X) is an X-tree (T, ϕ) with the requirement that ϕ is a bijective map from X to $L(T)$, the leaf set of T. Thus, in a phylogenetic tree, only the leaves are labeled, and each of them has exactly one label. For convenience, we sometimes refer to a leaf of a phylogenetic tree by its label. It is worth noting that there is no 2-degree vertex in a phylogenetic tree.

Theorem 5. *Let T be an undirected binary phylogenetic tree on X. Then T has $2|X| - 3$ edges and $|X| - 3$ internal edges.*

Rooted X-trees. A rooted X-tree is an ordered pair (X, ϕ), where T is a rooted tree and ϕ is a map from X to V in which every non-root vertex of degree at most two is labeled. Thus, in a rooted X-tree, the root is not necessarily labeled.

A rooted phylogenetic tree (on X) is a rooted X-tree (T, ϕ) in which ϕ is a bijective map from X to $L(T)$, and the root of T has degree at least 2.

Let T be a rooted phylogenetic tree with root ρ. If $\deg(\rho) \geq 3$, then we can consider T (with the direction of its edges ignored, if necessary) as an unrooted phylogenetic tree. If $\deg(\rho) = 2$, then we contract ρ (i.e. replace edges $\{\rho, v_1\}$ and $\{\rho, v_2\}$, where v_1 and v_2 are adjacent to ρ, by edge $\{v_1, v_2\}$), and consider the resulting graph as an unrooted phylogenetic tree.

Conversely, let T be an unrooted phylogenetic tree on X. Fix a particular element $x \in X$. We remove x, along with the pendant edge incident to it, and designate the vertex originally adjacent to x as the root of the resulting tree.

We have seen that there is a correspondence between rooted and unrooted phylogenetic trees. The following theorem formally characterizes this connection.

Let $P(X)$ and $R(X)$ be the set of unrooted and unrooted phylogenetic trees on X, respectively. Fix a particular element $x \in X$, and let $\psi_x \colon P(X) \to R(X \setminus \{x\})$ be a function that maps an unrooted tree $T \in P(X)$ to a rooted phylogenetic tree on $X \setminus \{x\}$ as described above.

Theorem 6. *The function ψ_x is bijective.*

We conclude the lecture with some additional notation. A cherry is a pair of leaves that have a common neighbor. A caterpillar (rooted or unrooted) is a binary phylogenetic tree for which the subgraph induced on the set of internal vertices is a path. A phylogenetic tree with no internal edge is called a star tree.

Let T be a rooted phylogenetic tree on X, and consider the edges of T as being directe away from T (thus, we can say vertices as being children or parents of others). The pendant subtree of T, denoted as $T(v)$, is the induced subgraph of T on the set of descendant vertices of v (including v itself). The cluster induced by v, denoted as $C_T(v)$, is the leaf-label set of $T(v)$.

Let X' be a subset of X. The least common ancestor of X' in T is a vertex $v \in V(T)$ with the property that: $C_T(v) \supset X'$; if $C_T(w) \supset X'$, where $w \neq v$, then w is a proper ancestor of v. The least common ancestor of X' is denoted by $\mathrm{lca}_T(X')$.

Let T be an unrooted phylogenetic tree on X. The removal of an edge e created exactly two connected components with leaf-label sets A and B, where $A \cup B = X$. We say that e induces a split $A|B$.

Finally, for specifying a rooted tree in the text format (e.g. as input to a program), we can use the Newick notation. In this notation, parentheses are used to group siblings, which are distinguished from each other by commas. For example, (a, b, c) represents a rooted star tree with three leaves; $((a, b), c), d)$ represents a rooted caterpillar in which (a, b) is the only chery. More details about the format can be found in http://evolution.genetics.washington.edu/phylip/newicktree.html.