Topic 4: Graphs

02-680: Essentials of Mathematics and Statistics

September 11, 2024

1 The Basics

Typically we denote a graph as the tuple $G = \langle V, E \rangle$, where V is a set of nodes or ***vertices*** and

$$E \subseteq V \times V \qquad \text{or} \qquad E \subseteq \{\{u, v\} : u, v \in V\}$$

is a set of ***edges*** or connections between two vertices. Which definition we use is dependent on G being ***directed*** or ***undirected*** (i.e. does the order of the nodes in the edge set matter).

Some examples of things that can be represented as graphs: train/flight maps, cell interactions, social networks, etc. In computational biology, we also represent genomes as graphs, in this case as a directed graph with changes denoted by different ***paths*** from start to finish. A *path* is a sequence of nodes

$$\langle v_1, v_2, v_3, ..., v_k \rangle$$

($k \geq 1$) where every edge

$$\langle v_i, v_{i+1} \rangle \in E \quad (\text{or } \{v_i, v_{i+1}\} \in E).$$

We say the path has ***length*** k, and that is a path *from* v_1 *to* v_k.

In the case of a genome graph, each edge would be ***labeled*** with a string, but in other types of graphs the labels could be numerical (in which we would usually call them *weights*). A path label would then be the concatenation of all the edge labels, a path weight would be the sum of the weights. We can also label or weight edges in some scenarios. In all cases weights (labels) are typically represented by a function:

$$\ell : E \to \Sigma^*$$

(or maybe $w : E \to \mathbb{R}$).

A lot of talked we want to talk about the ***neighborhood*** of a node:

$$N_G(v) = \{u \mid \{u, v\} \in E\}$$

and we call the size of this set

$$degree_G(v) := |N_G(v)| + \mathbb{1}\left(v \in N_G\left(v\right)\right)$$

the cardinality of the node or the ***degree***. Here $\mathbb{1} : \{\texttt{true}, \texttt{true}\} \mapsto \{1, 0\}$ in an indicator function that produces the value 1 when the input is true, and 0 otherwise. In a directed graph we can restrict this to an in- and out-degree (and neighborhood)

$$N_G^{in}(v) = \{u \mid \langle u, v\rangle \in E\}$$

$$N_G^{out}(v) = \{u \mid \langle v, u\rangle \in E\}$$

and thus $N_G(v) = N_G^{in}(v) \cup N_G^{out}$ (and the degrees are the respective set cardinalities). We call all of the nodes in $N_G(v)$ ***adjacent*** to v.

A graph is ***regular*** if

$$\forall v \in V : degree_G(v) = k$$

for some fixed k.

Note many times we will leave off the G and simply write $N, N^{in}, N^{out}, degree$ when the graph in question is clear from context.

Complete Graphs. A complete graph is one where all edges are present. That is for $G = \langle V, E\rangle$ to be complete,

$$E = V \times V \setminus \{\langle v, v\rangle \mid v \in V\} \qquad \text{or} \qquad E = \{\{u, v\} : u \neq v \in V\}$$

We also know in that case that it is $|V| - 1$ regular. We also call complete graphs ***Cliques***, especially in the context of subgraphs (below), and we do that so often we have a special notation for that: $\mathcal{K}_n$ (for a clique of size n).

Subgraph. Many times we only want to look at a particular part of a graph, think about say a regulatory network. (A regulatory network is a directed graph with nodes that are genes, and edges going from one gene to another if the source somehow impacts the expression of the sink.) If we run an experiment and get a list of genes that are differentially expressed in some condition (i.e. the expression level is different from the null/healthy case) we may want to try and intuit something looking only at those changes.

So we define a ***subgraph*** $G' = \langle V', E'\rangle$ where $V' \subseteq V$ and

$$E' \subseteq \left\{\langle u, v\rangle \mid u, v \in V' \wedge \langle u, v\rangle \in E\right\}.$$

That is we choose a subset of nodes, and the edges associated only with the chose nodes. In the example above, we almost always would want to look at the ***induced*** subgraph, which is basically the same thing but E' is *equal* to the set of edges above rather than a subset.

1.1 Bipartite Graphs

A ***Bipartite*** graph is $G = \langle (A \cup B), E \rangle$ where

$$E \subseteq \{\{u, v\} \mid u \in A, v \in B\}$$

(in directed graphs edges can go from A to B or B to A) and $A \cap B = \emptyset$. That is, its a graph where the nodes can be separated into two groups, and no edge exists within the group. These graphs come up a lot when doing some sort of assignment, in which case the actual assignment is a subgraph with all of the nodes but some subset of edges. An example could be assigning students to a peer advisor; maybe A is the set of first year students, and B is the second years. E is $A \times B$, and the goal is to find a subgraph such that

$$\forall v \in A : |N^{out}_{G'}(v)| = 1$$

(the out-degree of each node in the subgraph is 1).

Connected. A graph G is ***connected*** if the following holds:

$$\forall u, v \in V : \exists \langle u, z_1, z_2, ..., z_k, v \rangle \in V^* : \langle u, z_1 \rangle, \langle z_1, z_2 \rangle, ...\langle z_k, v \rangle \in E$$

That is, the graph is connected if for every pair of nodes, there is a path between them. We say that a ***connected component*** is a subgraph of G that is connected (note a graph that is already connected will only have one connected component).

2 Trees

A tree is a special type of graph that is fully connected and has exactly $|V| - 1$ edges.

One consequence of this is that there are no ***cycles***; a cycle is a special type of path $\langle v_1, v_2, ...v_{k-1}, v_1 \rangle$ $(k \geq 2)$ that ends at the same node it started at. In a tree, nodes with only one neighbor

$$Leaves(G) := \left\{ v \in V \;\middle|\; degree(v) = 1 \right\}$$

are called ***leaves***. Nodes that are not leaves $(V \backslash Leaves(G))$ are called ***internal nodes***.

Many times we will deal with ***rooted*** trees, though not always. A rooted tree is one that has a specific node $r \in V$ designated as the root. We call all of its neighbors it's ***children***,

their neighbors (excluding the root itself) are grandchildren to the root and children to the respective nodes and so on. In rooted trees we can also sometimes need to talk about ***levels***.

In a rooted tree we call all of the nodes below a specific node and the induced graph on them a ***subtree***, note this is slightly different from a sub*graph*, since its both restricted to rooted graphs and is definitional an induced subgraph from a specific node.

Regular Trees. Up to now there was no limit on the number of neighbors (children) a node could have, and in general this is true, but many times we want to know some general property of a tree. We use the term ***k-ary trees*** (more commonly we will talk about ***binary tree***, where k is 2). A k-ary tree is one where every internal node has degree $\forall v \in V, v$ is internal : $|N(v)| \leq k + 1$. It makes more sense in rooted trees, it means each internal node has at most k children. So in a rooted binary tree, a node has at most 2 children.

Phylogeny. In biology we often use trees to show relatedness of species/individuals/strains, and we call them ***phylogeny***. In this case we *usually* label the leaves with known examples (exigent species, sequenced strains, etc.) and internal nodes represent (probably unknown) ancestors. Many times we also add edge weights to these graphs that encode evolutionary distance. In phylogeny species that are more related have a common ancestor farther from the root (or one that is *lower*). So in the phylogeny of Humans, Chimpanzee, and Mouse; Human and Chimp would share an ancestor, then only meet Mouse at the root.

2.1 From Graphs to Trees

This will be discussed much more in 02-613, but the process of extracting a tree from a graph is called finding an ***spanning tree***. That is, for a graph $G = \langle V, E \rangle$ find a subgraph $T = \langle V, E' \rangle$ where:

(1) $E' \subseteq E$

(2) $|E| = |V| - 1$

(3) T is connected.

Useful References

Liben-Nowell, "Connecting Discrete Mathematics and Computer Science, 2e". §11.2-11.4