GRAPH THEORY

Graphs, undirected and (later) directed
We have seen in the last lecture an example of a Bayesian Network, a probabilistic graphical model used in machine learning. Such models rely on directed acyclic graphs (DAGs). Other models, called Markov Random Fields rely on undirected graphs.

The Facebook “friends” graph is an undirected graph. The Twitter “following” graph is directed. The Internet is an undirected graph, while the World Wide Web is a directed graph. A prerequisites (directed) graphs tells you in what order to take courses.

Find other interesting examples of graphs at the following links:
http://partneringresources.com/wp-content/uploads/Mayas-LinkedIn-Network.gif
http://www.biomedcentral.com/content/figures/1471-2180-11-234-1.jpg
http://www.ladamic.com/img/IngredientSubstitutes.jpg

We shall also discuss weighted graphs: graphs whose edges are labeled with numbers. For example, consider how to how to lay electric cables or water pipes to minimize cost. More generally, both edges and vertices can be labeled with additional information, not necessarily numbers. In many ways, labeled graphs are the most common tool used to build mathematical models for solving problems computationally.

The origins of graph theory are likely related to Euler’s “Bridges of Königsberg” problem, and to map coloring, which we shall visit later.

Definition 18.1 An undirected graph is a pair $G = (V, E)$ where $V$ is a finite nonempty set of vertices or nodes and $E \subseteq 2^V$ is a finite (possibly empty) set of edges consisting only of subsets of cardinality 2.

Let $e$ be an edge $e = \{u, v\}$. The vertices $u$ and $v$ are called the endpoints of $e$. We will denote by $u-v$ (equivalently, $v-u$) both the edge with endpoints $u$ and $v$ and the fact that there exists an edge between $u$ and $v$. We also say (sometimes)that the edge $u-v$ is incident to $u$ and to $v$.

Two vertices such that $u-v$ are called adjacent (or neighbors).

1. Two edges that share an endpoint, e.g., $u-v$ and $v-w$, are called adjacent in some textbooks and incident in others!
We shall omit “undirected” and just say “graph”. Later, we shall refer to directed graphs as “digraphs”.

Note that our definition of an edge as set of nodes of cardinality 2 precludes “loops” or “parallel edges”. Such features can be useful for certain modeling tasks and more general definitions for graphs exist (see multigraphs). The kind of undirected graphs we will work with are usually called simple graphs in some textbooks.

**Definition 18.2** The degree of a vertex, deg(u), is the number of neighbors of u (the number of vertices adjacent to u). A vertex of degree 0 is called isolated.

**Lemma 18.3 (The Handshaking Lemma)** The sum of degrees of all nodes in a graph is twice the number of edges.

\[ \sum_{v \in V} \deg(v) = 2|E| \]

**Proof:** Since each edge is incident to exactly two vertices, each edge contributes two to the sum of degrees of the vertices. ■

**Example 18.4** Prove that in any graph there are an even number of vertices of odd degree.

**Solution:** Let \( V_e \) and \( V_o \) be the set of vertices with even degree and the set of vertices with odd degree respectively in a graph \( G = (V,E) \). Then,

\[ \sum_{v \in V} \deg(v) = \sum_{v \in V_e} \deg(v) + \sum_{v \in V_o} \deg(v) \]

The first term on RHS is even since each vertex in \( V_e \) has an even degree. From the previous example, we know that LHS of the above equation is even. Thus, the second term on the RHS must be even. Since each vertex in \( V_o \) has odd degree, for the sum of the degrees of vertices in \( V_o \) to be even, \(|V_o|\) must be even. ■

**Definition 18.5** A graph in which every vertex is isolated is called edgeless.

The complete graph on \( n \geq 1 \) vertices, \(|V| = n\), notation \( K_n \), has as edges all the possible subsets of \( V \) of size 2, therefore \( \binom{n}{2} \) edges.

\( K_n \) has the maximum number of edges that are possible in a graph with \( n \) vertices, i.e., we always have

\[ |E| \leq \frac{|V|(|V| - 1)}{2} \]

Here are some other graphs of interest.
Definition 18.6 The path graph on \( n \geq 1 \) vertices, is \( P_n = ([1..n], \{k-(k+1) \mid k \in [1..(n-1)]\}) \). It has \( n \) vertices but only \( n-1 \) edges.

The cycle graph on \( n \geq 1 \) vertices, is \( C_n = ([1..n], \{k-(k+1) \mid k \in [1..(n-1)]\}) \cup \{n-1\} \). It has \( n \) vertices and \( n \) edges.

The \( m \times n \) grid has as vertices the pairs in \([1..m] \times [1..n]\) and as edges \((i,j)-(i+1,j)\) for each \( i \in [1..(m-1)]\) and each \( j \in [1..n]\) as well as \((i,j)-(i,j+1)\) for each \( i \in [1..m]\) and each \( j \in [1..(n-1)]\). It has \( mn \) vertices and \((m-1)n + m(n-1) = 2mn - m - n\) edges.

Definition 18.7 A walk is a non-empty sequence of vertices consecutively linked by edges: \( u_0, u_1, \ldots, u_k \) such that \( u_0 - u_1 - \cdots - u_k \). We call this a walk from \( u_0 \) to \( u_n \) (the endpoints of the walk) and we say that \( u_0 \) and \( u_k \) are connected by this walk. The length of this walk is the number \( k \) of edges (not the number \( k+1 \) of nodes). If \( u \) is a vertex, then \( u \) is also, by definition, a walk, whose length is 0. (Note that our graphs are not allowed to have edges linking a node with itself, but any node is connected to itself by a walk of length 0.

Definition 18.8 A path is a walk in which all vertices are distinct.\(^2\) Walks of length 0 are paths. Walks of length 1 are also paths because they consist of a single edge hence the endpoints are distinct.

Proposition 18.9 If \( u_0 - u_1 - \cdots - u_{n-1} - u_n \) is a walk of length \( n \geq 2 \) such that \( u_0 \neq u_n \), then there exist \( 1 \leq m \leq n-1 \) vertices \( v_1, \ldots, v_m \) such that \( u_0 - v_1 - \cdots - v_m - u_n \) is a path. (Where there is a walk, there is a path!)

Proof: This can also be shown by induction on the length of the walk but we shall give an elegant proof based on the Well-Ordering Principle (WOP).

Consider the walk \( u_0 - u_1 - \cdots - u_{n-1} - u_n \) of length \( n \geq 2 \) Define

\[
A = \{ k \mid k \geq 2 \text{ and there is a walk of length } k \text{ from } u_0 \text{ to } u_n \}
\]

\( A \) is non-empty because \( n \in A \). By WOP, \( A \) has a least element, call it \( p \). Note that \( 2 \leq p \leq n \).

So there must exist a walk of length \( p \) from \( u_0 \) to \( u_n \). Such a walk has \( p+1 \) vertices so it is of the form \( u_0 - v_1 - \cdots - v_m - u_n \) where \( m = p - 1 \).

We claim that this minimum length walk is a path and we prove this claim by contradiction. Suppose that two of its nodes coincide. This can happen in three different ways: \( u_0 \equiv v_i \) for some \( 1 \leq i \leq m \), or \( v_i \equiv v_j \) for some \( 1 \leq i < j \leq m \), or \( u_n \equiv v_i \) for some \( 1 \leq i \leq m \). In each case we can show that a strictly shorter walk exists between \( u_0 \) and \( u_n \), intuitively by “cutting off” cycles caused by repeated nodes. (more details in class).

\(^2\) Warning: some textbooks call “path” what we call “walk” and they call “simple path” what we call “path”.

3
Therefore, we can replace “walk” with “path” without loss of generality. This is useful because, in general, walks can be arbitrarily long but the length of paths is bounded by $|E|$ since edges cannot repeated (this would repeat vertices).

The statement of the proposition we just proved can “sharpened”, i.e., made stronger, to say that “every walk contains a path with the same endpoints”. This is true for walks of length 0 or 1. Beyond that, the precise statement is:

**Proposition 18.10** If $u_0-u_1-\cdots-u_{n-1}-u_n$ is a walk of length $n \geq 2$ such that $u_0 \neq u_n$, then there exist $1 \leq m \leq n-1$ vertices $v_1, \ldots, v_m$ such that $u_0-v_1-\cdots-v_m-u_n$ is a path whose sequence of nodes and edges is a subsequence of the sequence of nodes and edges of $u_0-u_1-\cdots-u_{n-1}-u_n$.

(Here, “subsequence” preserves order, but it does not necessarily consist of consecutive elements, i.e., $e_1e_3e_4$ is a subsequence of $e_1e_2e_3e_4$.)

The intuition behind this proposition is quite clear. We omit the proof but it is a good exercise for you to prove this by strong induction on the length of the walk.

### Equivalence relations

**Definition 18.11** A binary relation (sometimes relationship) on a set $A$ consist of pairs of elements of $A$ (therefore, it is a subset of $A \times A$). If $\rho$ is a binary relation on $A$ and $a, b \in A$, it is customary to write $a \rho b$ instead of $(a, b) \in \rho$.

A binary relation on a non-empty set $A$ is reflexive if for all $a \in A$ we have $a \rho a$, is symmetric if $a \rho b$ implies $b \rho a$, and is transitive if $a \rho b$ and $b \rho c$ imply $a \rho c$. A binary relation that is reflexive, symmetric, and transitive is called an equivalence relation.\(^3\)

**Proposition 18.12** Let $G = (V, E)$ be an undirected graph. We define a binary relation on $V$ denoted $u-\cdots-v$, read $u$ is connected to $v$, by the existence of a path/walk from $u$ to $v$.

Then, $-\cdots-$ is an equivalence relation on $V$.

**Proof:** To show reflexivity we use the path of length 0. To show transitivity we use the concatenation of two walks that share an endpoint. To show symmetry we use the reversal of a walk/path. \(\blacksquare\)

\(^3\)When $A$ is empty all these properties hold vacuously. We are not interested in this case and we have ruled it out by requiring $A \neq \emptyset$. 

4
Definition 18.13  An equivalence relation $\rho$ on $A$ defines certain special subsets of $A$. For each $a \in A$, the equivalence class of $a$ is

$$[a] = \{b \in A \mid b \rho a\}$$

Proposition 18.14  Let $\rho$ be an equivalence relation on $A$. For any $a, b \in A$ we have $[a] = [b]$ iff $a \rho b$.

Proof: By reflexivity, $a \in [a]$. Thus, if $[a] = [b]$ then $a \in [b]$ so $a \rho b$.

Conversely, if $a \rho b$ then for any $c \in [a]$ we have $c \rho a$, and, by transitivity, $c \rho b$ hence $c \in [b]$. We have just shown that $[a] \subseteq [b]$. Using symmetry and a similar argument we can show that $[b] \subseteq [a]$. ■

By this lemma, the equivalence classes partition $A$, i.e., they form a set of non-empty subsets $B_1, \ldots, B_m$ of $A$, called blocks that are pairwise disjoint and whose union is the entire $A$.

Conversely, given a partition $\{B_1, \ldots, B_m\}$ of a non-empty set $A$ define the following relation: $a \rho b$ when $a$ and $b$ are both in some block $B_i$.

Proposition 18.15  The relation $\rho$ is an equivalence relation.

Proof: Any $a \in A$ is in some $B_i$. Hence $apa$ so the relation is reflexive.

The relation is symmetric by the way it was defined.

Finally suppose $a \rho b$ and $b \rho c$. Then $a, b$ are both in some $B_i$ and $b, c$ are both in some $B_j$. Since $b$ is in both $B_i$ and $B_j$ we must have $B_i = B_j$. Hence $a \rho c$. So the relation is transitive. ■

Therefore we have shown how to associate a partition of $A$ to any equivalence relation on $A$. We have also shown how to associate an equivalence relation on $A$ to any partition of $A$. In fact, these two associations can be shown to be inverse to each other (no proof in these notes, but it’s a good exercise). Hence we have a one-to-one correspondence between the equivalence relations on $A$ and the partitions of $A$. In this precise sense, the concepts of equivalence relation and partition are “interchangeable”. [4]

Connectivity

Definition 18.16  Since it is an equivalence relation, the is-connected-to relation partitions $V$ into blocks of nodes mutually connected by paths. These blocks are called connected components. We will refer to the set of connected components as $CC$.

[4]To use a term beloved to Professor Gian-Carlo Rota, these two concepts are “cryptomorphic”. 5
There are graphs with just one connected component, e.g., $K_n$, $P_n$, $C_n$, grids. On the other hand, edgeless graphs have $|V|$ connected components, each of which consists of a single node. More generally, we have the following:

**Proposition 18.17** Every graph with $n \geq 1$ vertices and $m \geq 0$ edges has at least $\max(n - m, 1)$ and at most $n$ connected components. That is

$$\max(|V| - |E|, 1) \leq |CC| \leq |V|$$

**Proof:** The second inequality is immediate because each block of a partition must contain at least one element.

The max in the first inequality takes care of the cases when $n - m \leq 0$ because $V$ is nonempty and thus we have at least one connected component.

For the rest, we prove the proposition by induction on $m$.

**Base Case:** $m = 0$. As noted above, in the edgeless graph $|CC| = |V|$.

**Induction Step:** Let $k$ be an arbitrary natural number. Assume (IH) that every graph with $n$ vertices and $k$ edges has at least $n - k$ connected components.

We want to prove that a graph, $G$, with $n$ vertices and $k + 1$ edges has at least $n - (k + 1) = n - k - 1$ connected components. Note that $G$ has at least one edge. Consider the graph $G'$ that has the same nodes $G$ and is obtained by removing an edge (does not matter which one), say $\{u, v\}$, from $G$. The graph $G'$ has $n$ vertices and $k$ edges. By IH, $G'$ has at least $n - k$ connected components. Now add $\{u, v\}$ to $G'$ to obtain the graph $G$. We consider the following two cases.

**Case I:** $u$ and $v$ belong to the same connected component of $G'$. In this case, adding the edge $\{u, v\}$ to $G'$ is not going to change any connected components of $G'$. Hence, in this case the number of connected components of $G$ is the same as the number of connected components of $G'$ which is at least $n - k > n - k - 1$.

**Case II:** $u$ and $v$ belong to different connected components of $G'$. In this case, the two connected components containing $u$ and $v$ become one connected component in $G$. All other connected components in $G'$ remain unchanged. Thus, $G$ has one less connected component than $G'$. Hence, $G$ has at least $n - k - 1$ connected components.

It might be difficult to memorize the previous result because $|V| - |E|$ can be negative and therefore we needed the max to lower bound the number of connected components. Easier to remember, and equivalent to the core of the result we proved is the following inequality between two nonnegative integers. In any graph $G = (V, E)$ we have

$$|E| \geq |V| - |CC|$$

Observe that a set $C$ of nodes of $G$ is a connected component iff it is a *maximally connected* subset. i.e., any node of $G$ not in $C$ is not connected to any node in $C$. This follows immediately from the properties of equivalence relations.
Definition 18.18  A graph in which any two vertices are connected is called connected, otherwise it is called disconnected. A connected graph has exactly one connected component.

The graphs $K_n, P_n, C_n$ and the $m \times n$ grids are connected.

Example 18.19  Prove that every connected graph with $n$ vertices has at least $n - 1$ edges.

Solution: We will prove the contrapositive, i.e., a graph $G$ with $m \leq n - 2$ edges is disconnected. From Proposition 18.17 we know that the number of connected components of $G$ is at least $n - m$. But $n - m \geq n - (n - 2) = 2$. Hence there are at least 2 connected components and the graph is disconnected. 

Therefore, for every connected graph we have $|E| \geq |V| - 1$. 

7