BST 267: Introduction to Social and Biological Networks

Lecture 2

JP Onnela

Department of Biostatistics
Harvard T.H. Chan School of Public Health
Harvard University

October 26, 2022

Network Metrics and Algorithms I

Sets

- Intuitively speaking, a **set** is any collection of objects
- These objects are referred to as the **elements** of the set
- For example, $A = \{1, 2, 3\}$
- The order in which the elements of a set are listed is irrelevant
- We write $x \in A$ if x (whatever it may be) is an element of A
- We write $x \notin A$ if x is not an element of A
- Given two sets A and B, we say that A is a subset of B, denoted by $A \subseteq B$, if every element of A is also an element of B
  - For example, if $A = \{1, 2, 3\}$ and $B = \{1, 2, 3, 4, 5\}$, then A is a subset of B
- Set A is **equal** to set B if $A \subseteq B$ and $B \subseteq A$, i.e., A and B consist of exactly the same objects, in which case we write $A = B$

Vertices, Edges, and Graphs

- Graphs are mathematical representations of network structures
- A graph is a way of specifying relationships among a collection of items
- Graphs consist of two kinds of components:
  - Vertices (nodes)
  - Edges (ties, arcs)

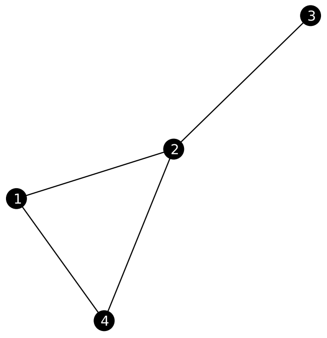
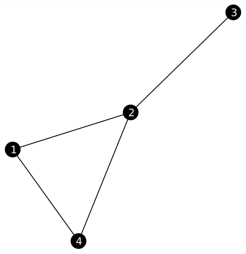


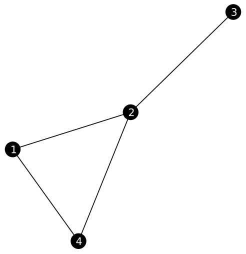
Figure: A graph of 4 nodes and 4 edges.

- A simple graph is an ordered pair $G = (V, E)$
- Here V (or $V(G)$) is the **vertex set** and E (or $E(G)$) is the **edge set** of graph G
- The vertex set here consists of vertices $V = \{1, 2, 3, 4\}$
- The edge set here consists of pairs of vertices $E = \{(1, 2), (1, 4), (2, 4), (2, 3)\}$
- The vertex pairs may be **ordered or unordered**, corresponding to directed and undirected graphs
- The edge set E can also be presented as an unordered list to encode the structure of a graph, in which case it is usually referred to as an **edge list**

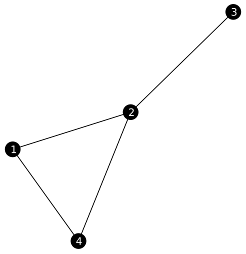


Vertices, Edges, and Graphs

- The graph here consists of four vertices labelled $1, 2, 3, 4$
- It is common, but not necessary, to label the vertices with numbers; we could have used the letters a, b, c, d instead
- Some vertex pairs are connected by an edge and some vertex pairs are not connected
- Two connected vertices are said to be (nearest) **neighbors**

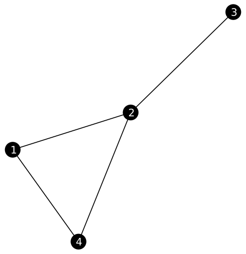


- Edges, depending on the context, can signify a variety of things
- Common interpretations
  - Structural connections
  - Interactions
  - Relationships
  - Dependencies
- Often more than one interpretation may be appropriate



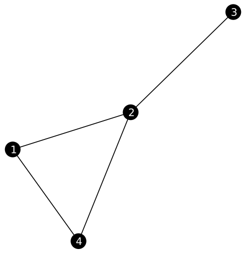
Vertex Degree

- The **degree** of a vertex in a graph is the number of edges connected to it
- We use k_i to denote the degree of vertex i
- Adopt standard notation for sums:
  $\sum_{i=m}^{n} x_i = x_m + x_{m+1} + x_{m+2} + \cdots + x_{n-1} + x_n$



Vertex Degree

- Every edge in an undirected graph has two “symmetric” ends
- If there are M edges in total, then there are $2M$ **ends of edges**
- The number of ends of edges is also equal to the sum of the degrees of all the vertices: $2M = \sum_{i=1}^{N} k_i$



Equal Graphs

- Consider two graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$
- Two graphs G_1 and G_2 are **equal** if they have equal vertex sets and equal edge sets, i.e., if $V_1 = V_2$ and $E_1 = E_2$
- Note that equality of graphs is defined in terms of equality of sets

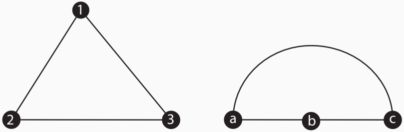


Figure: Are these two graphs equal?

Isomorphic Graphs

- Need a new concept of sameness
- Two graphs are **isomorphic** if there exists a one-to-one correspondence between their vertex sets with the property that whenever two vertices are adjacent in either graph, the corresponding two vertices are adjacent in the other graph
- If graphs G and H are isomorphic, we write $G \cong H$
- Isomorphism is a special one-to-one correspondence in that it not only associates vertices with vertices but also edges with edges

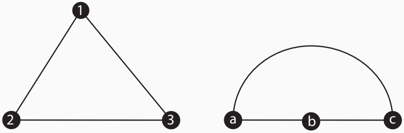
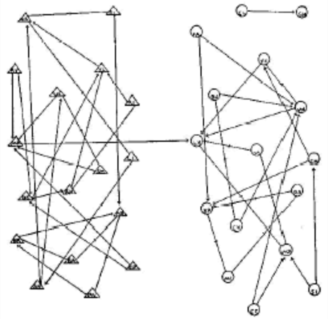


Figure: Two isomorphic graphs.

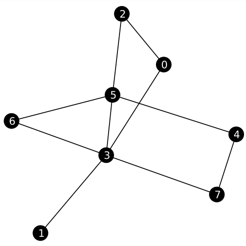
Subgraphs

- A graph H is a **subgraph** of G if $V(H) \subseteq V(G)$ and $E(H) \subseteq E(G)$
- Consider some subset of vertices $V'(G) \subseteq V(G)$; an **induced subgraph** of G is a subgraph $G' = (V', E')$ where $E(G') \subseteq E(G)$ is the collection of edges to be found in G among the subset $V(G')$ of vertices
- For example, consider Moreno's sociogram and let $V(G)$ represent all the vertices
- If we use V' to denote the set of vertices corresponding to boys, what is the graph G' induced by V'?

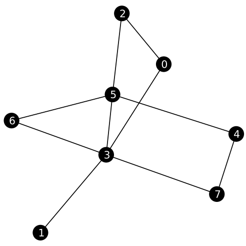


Walks, Trails, and Paths

- In mathematics, a **sequence** is an ordered list of objects, e.g., $(2, 4, 6)$
- A **walk** in a graph is a sequence $(v_1, v_2, v_3, \ldots, v_{n-1}, v_n)$ of not necessarily distinct vertices in which v_1 is joined by an edge to v_2, v_2 is joined by an edge to v_3, ..., v_{n-1} is joined by an edge to v_n
- A walk is sometimes presented as an alternating sequence of vertices and edges, such that every edge joins the vertices immediately preceding and following it; since the edges are obvious after we state the vertices, we use the simpler notation
- A walk $(v_1, v_2, v_3, \ldots, v_n)$ in a graph is a **closed walk** if v_1 and v_n are the same vertex; otherwise it is an **open walk**



- A **path** is a walk without repeated vertices
- A **trail** is a walk without repeated edges
- This means that every path is a trail, but not every trail is a path



- A vertex v in a graph is said to be **reachable** from another vertex u if there exists a path from u to v, i.e., if there is a way to get from u to v
- A graph is said to be **connected** if every vertex is reachable from every other vertex, i.e., if there is a path from every vertex to every other vertex

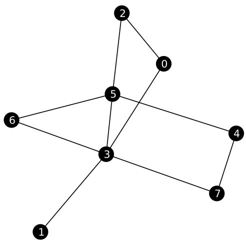


Figure: A connected graph.

Connectedness

- If a graph is not connected, it is said to be **disconnected**
- There is often no a priori reason to expect graphs to be connected

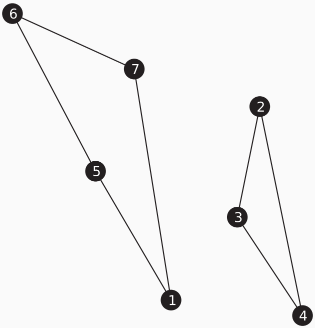
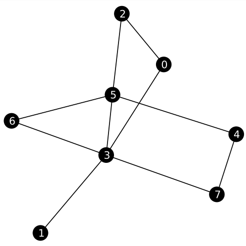


Figure: A disconnected graph.

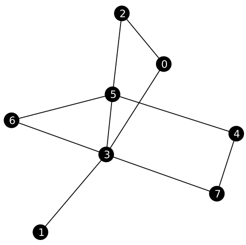
Path Lengths

- In addition to asking whether two nodes are connected by a path, it is interesting to ask how long such a path is (provided it exists)
- For example, the Internet is efficient at routing data because most routers are only a few hops from other routers (short paths); the same is true for diseases that spread via person-to-person contacts
- The **length of a path** is defined as the number of edges in the sequence that comprises it
- For example, the path $(3, 6, 5, 2)$ in the graph below consists of the edges $((3, 6), (6, 5), (5, 2))$ and therefore has length three



Path Lengths

- We can use path lengths to quantify distance between two nodes in a graph
- This leads us to consider the **shortest path** (or, possibly, paths) connecting any given two nodes
- The **distance** between vertex u and vertex v is defined as the length of the shortest path between them
- For example, there are two equally short paths between vertices 3 and 2, which are $(3, 5, 2)$ and $(3, 0, 2)$, both of which have a length of 2
- **Diameter** is defined as the length of the longest of all pairwise shortest paths
- What is the diameter of the graph below?



Link Density

- Consider an undirected network with N nodes
- Recall that an edge is an (here, unordered) vertex pair
- How many edges can the network have at most?
- The number of possible edges is equal to the number of ways of choosing 2 vertices out of N

$$\binom{N}{2} = \frac{N!}{(N-2)!2!} = \frac{N(N-1)}{2} \tag{1}$$

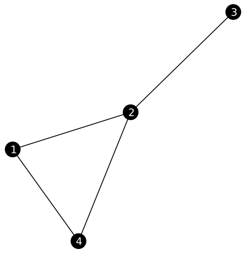
- How can we reason this without combinatorics?
- A graph is said to be **fully connected** if all possible edges are present

Link Density

- Let the number of edges be L
- The fraction of links present is called **link density** and is denoted by d (or ρ):

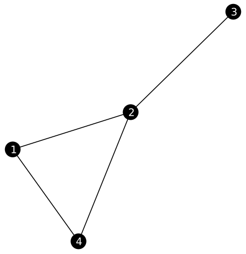
$$d = \frac{L}{N(N-1)/2} \tag{2}$$

- Link density by construction lies in the $[0, 1]$ interval
- Most networks have very low values of density



Link Density

- Networks generated with models can be said to be dense or sparse
- The concept does not refer to a specific value of d
- Instead, we need to consider a network growth process and ask what happens as the number of nodes $N \to \infty$
  - If d tends to a constant as $N \to \infty$ the network is said to be **dense**
  - If d tends to zero as $N \to \infty$ the network is said to be **sparse**



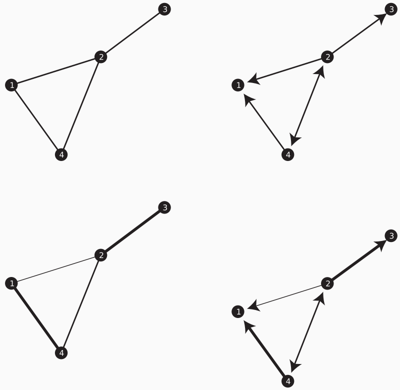


Figure: Different types of graphs.

There are many different types of graphs:

- Simple graphs
  (unweighted, undirected, symmetric)
- Directed graphs
  (unweighted, asymmetric)
- Weighted graphs
  (undirected, symmetric)
- Weighted and directed graphs
  (asymmetric)

- An undirected graph is represented by an $N \times N$ (symmetric) adjacency matrix $\mathbf{A}$

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \cdot & \cdot & \cdots & \cdot \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{pmatrix} \tag{3}$$

- For a simple (unweighted, undirected, symmetric) graph

$$A_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

- For an undirected graph of N vertices, the degree can be written in terms of the adjacency matrix as $k_i = \sum_{j=1}^{N} A_{ij}$

Representing Networks

- The **transpose** $\mathbf{A}^{\mathrm{T}}$ of an $N \times N$ matrix $\mathbf{A}$ is the $N \times N$ matrix that has the first row of $\mathbf{A}$ as its first column, the second row of $\mathbf{A}$ as its second column, etc.
- A matrix is said to be **symmetric** if $\mathbf{A}^{\mathrm{T}} = \mathbf{A}$
- The adjacency matrices of undirected graphs are always symmetric, whereas for directed graphs generally $\mathbf{A} \neq \mathbf{A}^{\mathrm{T}}$
- In statistics $\mathbf{A}$ is sometimes replaced with the matrix $\mathbf{X}$ with elements X_{ij}

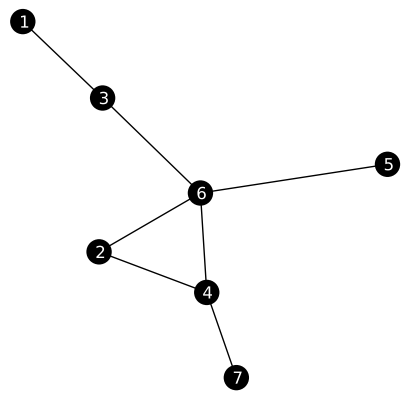
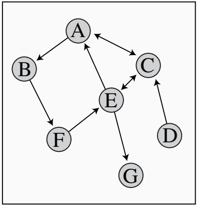


Figure: Example of a simple graph.

$$\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

- Directed graphs are called **digraphs** for short
- The adjacency matrix of a directed graph has element $A_{ij} = 1$ if there is an edge **from vertex** i **to vertex** j (convention)
- The adjacency matrices associated with digraphs are usually not symmetric



$\longrightarrow$

	A	B	C	D	E	F	G
A	0	1	1	0	0	0	0
B	0	0	0	0	0	1	0
C	1	0	0	0	1	0	0
D	0	0	1	0	0	0	0
E	1	0	1	0	0	0	1
F	0	0	0	0	1	0	0
G	0	0	0	0	0	0	0

Figure: Graphical and matrix representation of a directed graph.

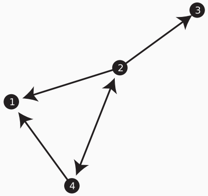
Vertex Degree

- In a directed network, each vertex has two degrees:
  - The **in-degree** is the number of incoming edges
  - The **out-degree** is the number of outgoing edges
- We can write in-degree of node j and out-degree of node i as:

$$
\begin{aligned}
k_j^{\text{in}} &= \sum_{i=1}^{N} A_{ij} \\
k_i^{\text{out}} &= \sum_{j=1}^{N} A_{ij}
\end{aligned}
$$

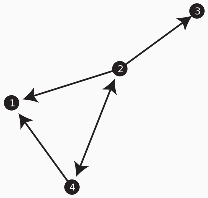
- Alternatively, we can write in-degree and out-degree of node i as:

$$
\begin{aligned}
k_i^{\text{in}} &= \sum_{j=1}^{N} A_{ji} \\
k_i^{\text{out}} &= \sum_{j=1}^{N} A_{ij}
\end{aligned}
$$



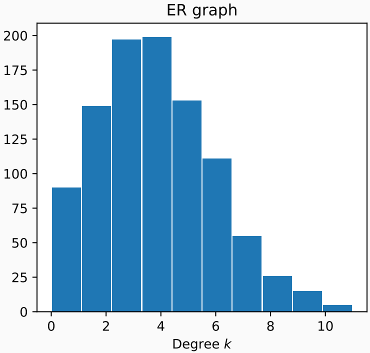
Vertex Degree

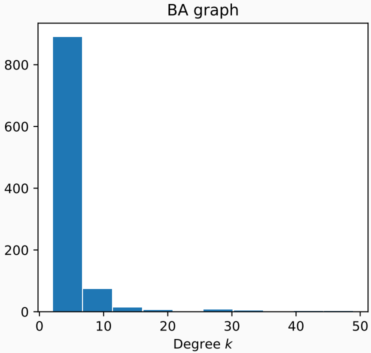
- Social science literature sometimes refers to in-degree as **popularity** and out-degree as **expansiveness**
- Statistical literature on networks sometimes uses a short-hand notation for sums:
  - In-degree: $k_j^{\mathrm{in}} = \sum_{i=1}^{N} A_{ij} = A_{+j}$ (row sum)
  - Out-degree: $k_i^{\mathrm{out}} = \sum_{j=1}^{N} A_{ij} = A_{i+}$ (column sum)



Vertex Degree

- The distribution of vertex degrees in a given graph is called the **degree distribution** of the graph
- Degree distribution is probably the single most important metric or description of any graph





Clustering Coefficient

- We often want to know how densely the neighbors of a given node are connected
- Consider a node i with degree k_i
- Let t_i denote the number of ties that exist among the neighbors of i
- **Local clustering coefficient** is defined as the number of ties that exist between the neighbors of i divided by the number of ties that could exist
- This gives rise to

$$c_i = \frac{t_i}{k_i(k_i - 1)/2} \tag{5}$$

- The mean local clustering coefficient in a network is computed by taking the mean of c_i over all nodes i in the network

Reciprocity

- Triangles are the shortest possible loop in an undirected network
- In directed networks, the shortest loop has length two with edges (i, j) and (j, i)
- We say that the edge (i, j) is reciprocated by the edge (j, i) (and vice versa)
- In a directed graph, the frequency of loops of length two is measured by **reciprocity**, which is defined as the fraction of edges that are reciprocated
- If there are a total of L directed edges in the network and L_m of them are mutual (reciprocated), then reciprocity is given by $r = L_m/L$
- Would we expect social ties or WWW links to be reciprocated?

Reciprocity

- Reciprocity can also be interpreted as the probability for the edge (j, i) to exist given that edge (i, j) exists
- For example, about 57% of web links are reciprocated
- Reciprocity can be computed using properties of the adjacency matrix $\mathbf{A}$
- A pair of nodes, connected or not, is called a dyad (pair of nodes)
- For the (i, j) dyad, the associated adjacency matrix elements are A_{ij} and A_{ji}
- The product of the elements $A_{ij}A_{ji}$ is 1 if and only if $A_{ij} = 1$ and $A_{ji} = 1$
- We can now write

$$r = \frac{1}{L}\sum_{i,j} A_{ij}A_{ji} \left(= \frac{1}{L}\mathrm{Tr}\mathbf{A}^2\right) \tag{6}$$

- Example of reciprocity $r = L_m/L$