Random graphs, social networks and the internet

Lecture 1

Mariana Olvera-Cravioto

UNC Chapel Hill

molvera@unc.edu

September 23rd, 2020

Social networks and graphs

- The internet, the web, Facebook, Twitter, LinkedIn, Instagram, WhatsApp, WeChat, Snapchat, Pinterest, Reddit, etc. are all examples of **networks**.
- In **social networks**, connections occur among people.
- A connection between two people can mean many different things depending on the network, e.g., friendship, hyperlinks, follower-followed relations, etc.
- There are also many networks that do not involve people at all, e.g., the internet, neural connections in the brain, interactions between proteins in biology, articles in a citation network, etc.
- When analyzing networks, it is often convenient to think of them as **graphs**.

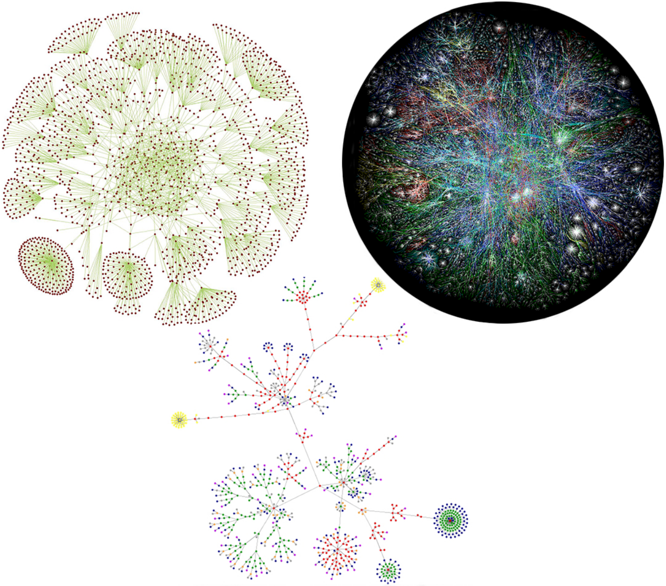
Graphs

- A graph consists on a set of **vertices**, V, and a set of **edges** E.
- Graphs can be **undirected** or **directed**.
- In an undirected graph, the relation between the vertices is symmetric, while in a directed graph it is not.
- We will call the vertices $V = \{1, 2, \ldots, n\}$, and write $i \to j$ to mean there is an edge (perhaps undirected) from vertex i to vertex j.
- In an undirected graph, the **degree** of a vertex is the number of edges incident to it.
- In a directed graph, the **in-degree** is the number of inbound edges and the **out-degree** is the number of outbound edges.

Examples of graphs

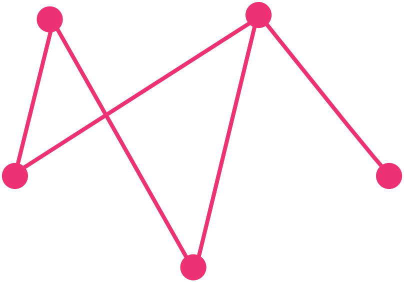
- **Internet:** vertices are servers/computers and edges are connections between them.
- **Web:** vertices are webpages and edges are directed links from one page to another.
- **Facebook:** vertices are people and edges are friendship relations.
- **Twitter:** vertices are people and edges go from followers to followed.
- **Citation network:** vertices are research articles and a directed edge represents a citation.

Different types of graphs



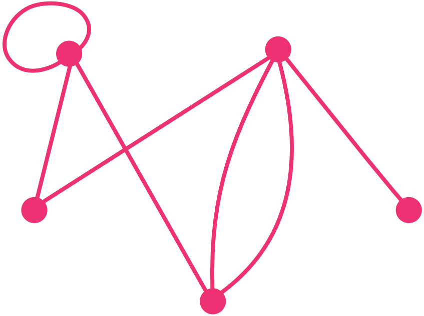
Types of graphs

- **Simple graphs:** a graph that has no self-loops nor multiple edges between any two vertices.



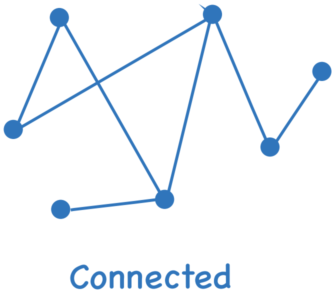
Types of graphs

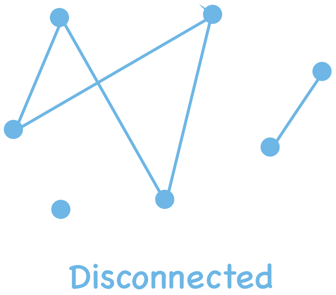
- **Multigraphs:** a graph that may have self-loops or multiple edges between two vertices.



Types of graphs

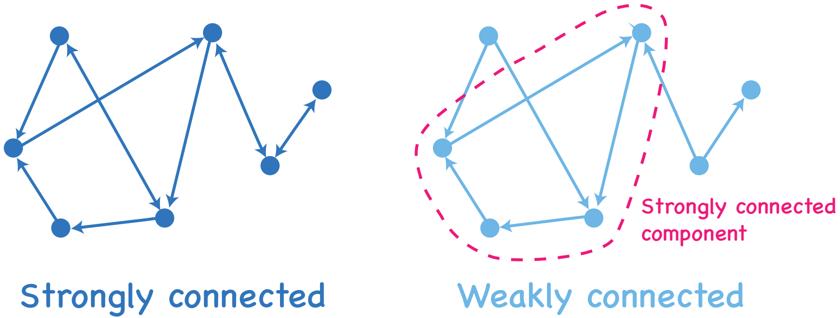
- **Connected graphs:** graphs where every pair of vertices is connected through a path.





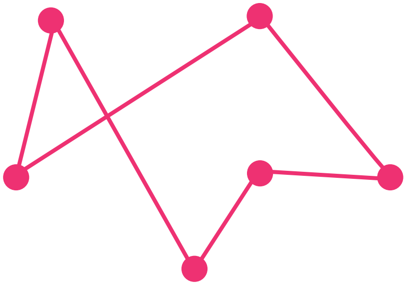
Types of graphs

- **Strongly connected graphs:** for directed graphs and any pair of vertices i and j, there exists a directed path from i to j and one from j to i.



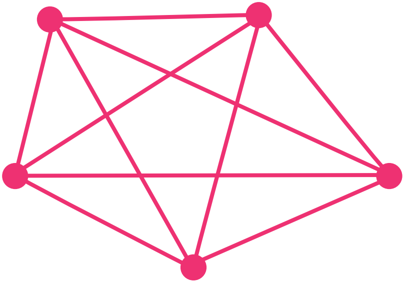
Types of graphs

- **Regular graphs:** all the vertices in the graph have the same degree.



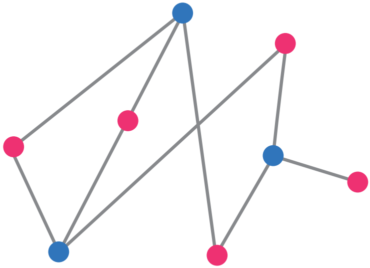
Types of graphs

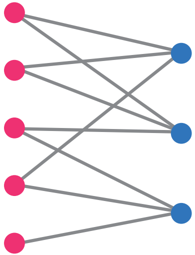
- **Complete graphs:** there is an edge between every pair of vertices in the graph.



Types of graphs

- **Bipartite graphs:** there are two classes of vertices, say V_1 and V_2, and edges occur only between a vertex in V_1 and one in V_2.





Structures and properties

- Some structures that can be of interest when studying graphs are:
  - **Cycles:** paths that start and end with the same vertex without repeated vertices.
  - **Cliques:** complete subgraphs.
  - **Distance between two vertices:** length of the minimum path connecting two vertices; in directed graphs the path must be directed.
  - **Component of a vertex:** the set of vertices that can be reached through (directed) paths from a given vertex.

- Some properties of interest:
  - **Diameter:** the maximum distance between two points in the graph.
  - **Components:** sizes of the largest, second largest, etc.
  - **Cycle lengths:** the typical length of cycles in the graph.
  - **Clustering:** the proportion of triangles (3-cliques) vs. open wedges.
  - **Communities:** subsets of vertices that have more edges among their vertices than with vertices outside the set.

Some questions of interest

- Is the graph (strongly) connected?
  - If not, does there exist a giant (strongly) connected component?
  - What is the size of the smaller components?
- What is the diameter of the graph?
- What is the typical distance between vertices in the graph?
- What is the degree distribution in the graph?
- Does the graph have clusters/communities?
- Are there vertices that are more “influential” or “central” to the network?

The small world phenomenon

- In the late 60's, a social psychologist named Stanley Milgram conducted a set of experiments to try to determine the typical length of paths connecting two individuals in the United States.
- A letter addressed to somebody in Boston would be given to a set of randomly chosen people in different states in the Midwest, strangers to the recipient, with the instruction to help it reach its destination by sending it to an acquaintance.
- **Result:** it took an average of 6 people to connect the first sender and the final recipient, something that became known as the

  small world or **six degrees of separation**

  phenomenon.
- Interestingly, the small world property is very common in large real-world networks.

Scale-free networks

- Recall that the degree of a vertex $i \in V = \{1, 2, \ldots, n\}$ in an undirected graph, denoted D_i, is the number of edges incident to it.
- The proportion of vertices having degree $k = 0, 1, 2, \ldots,$ is given by

$$p(k) = \frac{1}{n} \sum_{i=1}^{n} 1(D_i = k)$$

- We call $\{p(k) : k \geq 0\}$ the **degree distribution**.
- If the degree distribution of a graph satisfies

$$p(k) \propto k^{-\gamma}$$

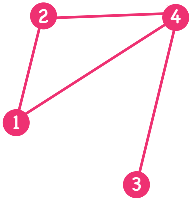
  for some $\gamma > 0$ (usually $\gamma \in (2, 3)$), we say that the graph is **scale-free**.
- In a scale-free graph there are vertices that have really large degrees, even if the average degree is small.

The adjacency matrix

- A convenient way to represent a graph is through its adjacency matrix.
- For a graph $G(V, E)$ having vertices $V = \{1, 2, \ldots, n\}$, its adjacency matrix A is the $n \times n$ matrix which has:

$$a_{i,j} = \text{number of edges from vertex } i \text{ to vertex } j$$

- In a **simple** graph we have $a_{i,j} \in \{0, 1\}$ for all (i, j).
- In a **undirected** graph the matrix A is symmetric.
- Example:



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

Counting paths

- Suppose A is the adjacency matrix of a graph $G(V, E)$.
- Then, for any $k \geq 1$ and any $i, j \in V$,

$$(A^k)_{i,j}$$

is the **number of paths of length k** from vertex i to vertex j, where A^k is the matrix A raised to the kth power.

Random graph models

- Some real networks are too big to be analyzed exactly.
- Some may even be constantly changing.
- **Idea:** we can think of our specific real-world graph as just one "typical" element of a larger class.
- If we can show that a property holds for a large class of graphs, it is likely it will hold for our specific graph.
- **Random graphs** are mathematical models that can help us understand large real-world graphs.
- No random graph model can mimic all the properties of a specific real-world graph, so we focus on choosing models that share certain properties that are important to the problem we want to analyze.

Large graph limit

- Random graph models consist of a vertex set $V_n = \{1, 2, \ldots, n\}$ and a set of rules for determining whether a given edge is present or not based on some random events.
- Their mathematical analysis is usually done under the **large graph limit** $n \to \infty$ on a sequence of graphs $\{G(V_n, E_n) : n \geq 1\}$.
- Taking the limit $n \to \infty$ simplifies computations in order for us to identify general properties.
- In practice, establishing results in the large graph limit means that our findings are likely to be true for sufficiently large graphs.

Static vs. evolving models

- Random graph models can be broadly classified into two categories: **static models** and **evolving or growing models**.
- Static models are meant to represent a "snapshot" of a large network.
- In static models $G(V_n, E_n)$ and $G(V_{n+1}, E_{n+1})$ can be totally different.
- Evolving models are meant to describe the growth of a graph as vertices get added to the graph (usually one at a time), so $G(V_n, E_n)$ and $G(V_{n+1}, E_{n+1})$ share most edges.
- In many evolving models edges and vertices never disappear, so $G(V_n, E_n)$ is a subgraph of $G(V_{n+1}, E_{n+1})$.

The Erdős-Rényi random graph

- The simplest model for a random graph is the **Erdős-Rényi model**.
- Consider a graph with vertex set $V_n = \{1, 2, \ldots, n\}$.
- There are a total of $\binom{n}{2}$ possible edges in the graph, and each of them will be chosen to be present or not with a coin flip.
- Suppose you have a coin that lands heads with probability $p \in (0, 1)$.
- For each pair of vertices i and j, toss the coin; if it lands heads, draw an edge between i and j, otherwise do nothing.
- Equivalently, if A denotes the adjacency matrix of the graph, let

$$a_{i,j} = a_{j,i} = 1(\text{coin-flip is a head}), \qquad i \neq j,$$

and set $a_{i,i} = 0$.

Properties of the Erdős-Rényi model

- This is the most studied random graph model there is.
- Some of its connectivity properties are:
  - If $np < 1$ the graph will consists of only small components of size $O(\log n)$.
  - If $np \to c > 1$ the graph will contain a unique *giant* connected component, with all other components of size $O(\log n)$.
  - If $np = 1$ the largest component will have size $O(n^{2/3})$.
  - If $p < (1-\epsilon)n^{-1}\log n$ the graph will most likely be **disconnected**.
  - If $p > (1+\epsilon)n^{-1}\log n$ the graph will most likely be **connected**.
- When the graph is connected, it exhibits the **small-world** property, with typical distance of order $O(\log n)$.

Degree distribution

- To compute the degree distribution we can use binomial probabilities.
- Fix a vertex $i \in V_n$, then its degree is given by

$$D_i = \sum_{j=1}^{n} \chi_{i,j}, \qquad \chi_{i,j} = 1((i,j) \in E_n)$$

- Note that the $\chi_{i,j}$ are independent Bernoulli r.v.s with parameter p.
- Therefore, since all vertices have the same distribution, for all $i \in V_n$,

$$P(D_i = k) = P(D_1 = k) = P(\text{Binomial}(n,p) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Moreover, if $np \to c$ as $n \to \infty$, we have that

$$\lim_{n\to\infty} P(D_1 = k) = \frac{e^{-c} c^k}{k!}, \qquad k \geq 0,$$

known as a Poisson distribution with mean c.... not **scale-free**.

Thank you for your attention.