WARWICK
THE UNIVERSITY OF WARWICK

Introduction to Graph Theory

Susana Gomes

Last updated November 18, 2023

Contents

1. Some properties and examples of graphs
2. Subgraphs and clustering
3. Graph Spectra and other centrality measures
4. Graph Laplacian

Networks - basic definitions and characteristics

Graphs - definition

Everything we will learn about networks is based on a concept which most of you have probably seen before - a graph. Today we will review basic definitions and concepts of graph theory which will be useful in the next few weeks.

Definition:

A **graph** (or **network**) $G = (V, E)$ consists of a finite set $V = \{1, \ldots, N\}$ of **vertices** (or **nodes**, **points**), and a set $E \subset V \times V$ of **edges** (or **links**, **lines**).

The graph G is called **undirected** if $(i, j) \in E$ implies $(j, i) \in E$, otherwise we say it is **directed**.

Two nodes $i, j \in V$ are **adjacent** or **neighbouring** if $(i, j) \in E$.

The structure of the graph is encoded in the **adjacency** (or **connectivity**) **matrix** which is defined as

$$A = (a_{ij} : i, j \in V) \quad \text{where} \quad a_{ij} = \begin{cases} 1 & , (i, j) \in E \\ 0 & , (i, j) \notin E \end{cases} .$$

We denote the number of edges by $K = |E|$ for directed, or $K = |E|/2$ for undirected graphs.

Some things to note about graphs

The definition of a graph is quite general, and with the way it is written in the previous slide we could allow for *self-edges*, i.e., edges like (i, i), or *multiple edges* (multiple instances of (i, j). We do not allow for this in this module.

One can also consider **weighted graphs**, which are graphs with edge weights $w_{ij} \in \mathbb{R}$. These can be used to represent continuous- or discrete-time Markov chains.

In general graphs can also be infinite, but we will focus on finite graphs. Many of the following graph characteristics only make sense in the finite case.

Graphs - paths and shortest paths

A **path** γ_{ij} of length $l = |\gamma_{ij}|$ from vertex i to j is sequence of vertices

$$\gamma_{ij} = (v_1 = i, v_2, \dots, v_{l+1} = j) \quad \text{with} \quad (v_k, v_{k+1}) \in E \text{ for all } k = 1, \dots, l,$$

and $v_k \neq v_{k'}$ for all $k \neq k' \in \{1, \dots, l\}$ (i.e. each vertex is visited only once).

If such a path exists, we say that vertex i is **connected** to j (write $i \to j$).

A **cycle** is a closed path γ_{ii} of length $|\gamma_{ii}| > 2$.

Shortest paths between vertices i, j are called **geodesics**. They are not necessarily unique, and their length d_{ij} is called the **distance** from i to j.

If $i \not\to j$ we set $d_{ij} = \infty$.

Graphs - connectivity

We say that a graph is **connected** if $d_{ij} < \infty$ for all $i, j \in V$.

We can define the **diameter** of the graph G by

$$\mathrm{diam}(G) := \max\{d_{ij} : i, j \in V\} \in \mathbb{N}_0 \cup \{\infty\},$$

and the **characteristic path length** of the graph G by

$$L = L(G) := \frac{1}{N(N-1)} \sum_{i,j \in V} d_{ij} \in [0, \infty].$$

Undirected graphs must have $d_{ij} = d_{ji}$ (which is finite if $i \leftrightarrow j$), and they can be decomposed into **connected components**, where we write

$$C_i = \{j \in V : j \leftrightarrow i\} \quad \text{for the component containing vertex } i.$$

Degrees

An important characteristic of any graph is the degree.

Definition:

The **in-** and **out-degree** of a node $i \in V$ is defined as

$$k_i^{\text{in}} = \sum_{j \in V} a_{ji} \quad \text{and} \quad k_i^{\text{out}} = \sum_{j \in V} a_{ij}.$$

$k_1^{\text{in}}, \dots k_N^{\text{in}}$ is called the **in-degree sequence**.

Using the in-degree sequence, we define the **in-degree distribution**:

$$\big(p^{\text{in}}(k) : k \in \{0, \dots, K\}\big) \quad \text{with} \quad p^{\text{in}}(k) = \frac{1}{N} \sum_{i \in V} \delta_{k, k_i^{\text{in}}},$$

which gives the fraction of vertices with in-degree k. [The same holds for out-degrees.]

In undirected networks, we simply write $k_i = k_i^{\text{in}} = k_i^{\text{out}}$ and $p(k)$.

Some properties of the degree

Here are some things to note about the degree of a graph:

- We always have $\sum_{i \in V} k_i = \sum_{i,j \in V} a_{ij} = |E|$.
- We can compute the average degree and the variance:

$$\langle k \rangle = \frac{1}{N} \sum_{i \in V} k_i = \frac{|E|}{N} = \sum_k k p(k), \qquad \text{and} \qquad \sigma^2 = \langle k^2 \rangle - \langle k \rangle^2.$$

- If a graph is such that each vertex has the same degree $k_i \equiv k$, we call it a **regular graph** (and it is usually undirected).
- Graphs where the degree distribution $p(k)$ shows a power law decay, i.e. $p(k) \propto k^{-\alpha}$ for large k, are often called **scale-free**.
  - ⋆ The majority of networks in applications has long-tailed degree distributions and in many cases this is given by a power law. α is typically between 2 and 3.
- Real-world networks are often scale-free with exponent around $\alpha \approx 3$.

Degree distribution

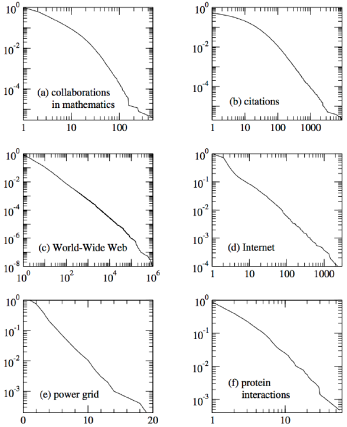

FIG. 9. Cumulative degree distributions for six different networks. The horizontal axis for each panel is vertex degree k (or in-degree for the citation and Web networks, which are directed) and the vertical axis is the cumulative probability distribution of degrees, i.e., the fraction of vertices that have degree greater than or equal to k. The networks shown are: (a) a collaboration network of mathematicians; (b) citations between 1981 and 1997 to all papers cataloged by the Institute for Scientific Information; (c) a 300 million vertex subset of the World Wide Web, *circa* 1999; (d) the Internet at the level of autonomous systems, April 1999; (e) the power grid of the western United States; (f) the interaction network of proteins in the metabolism of the yeast. Of these networks, three of them, (c), (d) and (f), appear to have power-law degree distributions, as indicated by their approximately straight-line forms on the doubly logarithmic scales. Taken from Newman, Mark EJ. "The structure and function of complex networks." SIAM review 45.2 (2003): 167-256.

(figure taken from the Networks lecture notes by R. Lambiotte, Oxford)

Graphs - first examples

Here are some examples of graphs:

- The **complete graph** K_N with N vertices is an undirected graph where all $N(N-1)/2$ vertices $E = ((i,j) : i \neq j \in V)$ are present.

- **Regular lattices** $\mathbb{Z}^d$ with edges between nearest neighbours are examples of regular graphs with degree $k = 2d$.

A special example

A particularly useful example of a graph is a tree:

Definition:

A **tree** is an undirected graph where any two vertices are connected by exactly one path.

In a tree, vertices with degree 1 are called **leaves**.

A **rooted tree** is a tree in which one vertex $i \in V$ is the designated **root**.

Then the graph can be directed, where all vertices point towards or away from the root.

Trees and cycles

Recall that a **cycle** is a closed path γ_{ii} of length $|\gamma_{ii}| > 2$. Using this, we can see that G is a tree if and only if

1. it is connected and has no cycles;
2. it is connected but is not connected if a single edge is removed;
3. it has no cycles but a cycle is formed if any edge is added.

Graphs - paths and cycles

A **path** γ_{ij} of length $l = |\gamma_{ij}|$ from vertex i to j is sequence of vertices

$$\gamma_{ij} = (v_1 = i, v_2, \ldots, v_{l+1} = j) \quad \text{with} \quad (v_k, v_{k+1}) \in E \text{ for all } k = 1, \ldots, l,$$

and $v_k \neq v_{k'}$ for all $k \neq k' \in \{1, \ldots, l\}$ (i.e. each vertex is visited only once).

If such a path exists, we say that vertex i is **connected** to j (write $i \to j$).

A **cycle** is a closed path γ_{ii} of length $|\gamma_{ii}| > 2$.

Shortest paths

Shortest paths between vertices i, j are called **geodesics**. They are not necessarily unique, and their length d_{ij} is called the **distance** from i to j.

Note that:

- For undirected graphs, d_{ij} is actually a distance.
- If G is a directed graph, you might have $d_{ij} \neq d_{ji}$!
- We can usually compute d_{ij} using Dijkstra's algorithm (not covered in this module)
- If $i \not\to j$ we set $d_{ij} = \infty$.

Connectivity

We say that a graph is **connected** if $d_{ij} < \infty$ for all $i, j \in V$.

Undirected graphs have $d_{ij} = d_{ji}$ (which is finite if $i \leftrightarrow j$), and they can be decomposed into **connected components**, where we write

$$C_i = \{j \in V : j \leftrightarrow i\} \quad \text{for the component containing vertex } i.$$

We can also define

- the **diameter** of the graph G by

$$\mathrm{diam}(G) := \max\{d_{ij} : i, j \in V\} \in \mathbb{N}_0 \cup \{\infty\},$$

- the **characteristic path length** of the graph G by

$$L = L(G) := \frac{1}{N(N-1)} \sum_{i,j \in V} d_{ij} \in [0, \infty].$$

⋆ Note that in many real networks, this value is really small compared to the number of nodes.

⋆ For example (taken from Lambiotte's lecture notes), a Facebook network with $N \approx 7.2 \times 10^8$ users (nodes) with 6.9×10^10 friendships (edges) yielded $L \approx 4.7$.

Degree correlations

Once we know the degrees of vertices of a graph and their distribution, we can start looking at some interesting properties.

In undirected graphs, we say that the **joint degree distribution** of two nodes linked by an edge is

$$q(k,k') = \frac{1}{|E|} \sum_{(i,j)\in E} \delta_{k_i,k}\delta_{k_j,k'} = \frac{\sum_{i,j\in V} a_{ij}\delta_{k_i,k}\delta_{k_j,k'}}{\sum_{i,j\in V} a_{ij}} = q(k',k).$$

We define the the marginal $q(k') = \sum_k q(k,k')$ and with it the **conditional degree distribution**

$$q(k|k') = \frac{q(k,k')}{q(k')} \quad \text{with average} \quad k_{nn}(k') := \sum_k kq(k|k').$$

The network is called

- **uncorrelated** if $k_{nn}(k')$ is independent of k',
- **assortative** if $k_{nn}(k') \nearrow$ in k',
- **disassortative** if $k_{nn}(k') \searrow$ in k'.

More about degree correlations

The marginal $q(k)$ corresponds to **edge sampled degree distribution**:

$$q(k) = \sum_{k'} q(k,k') = \frac{1}{|E|} \sum_{i,j \in V} a_{ij} \delta_{k_i,k} = \frac{N}{|E|} \frac{1}{N} \sum_{i \in V} k_i \delta_{k_i,k} = \frac{kp(k)}{\langle k \rangle} \ .$$

For uncorrelated networks we have $q(k|k') = q(k)$ and so $\quad k_{nn}(k') = \frac{\langle k^2 \rangle}{\langle k \rangle}$.
Note that this is not necessarily the same as the mean degree!

The degree of correlation can be quantified by the **correlation coefficient**

$$\chi := \frac{\langle kk' \rangle_q - \langle k \rangle_q^2}{\langle k^2 \rangle_q - \langle k \rangle_q^2} = \frac{\sum_{k,k'} kk' \left(q(k,k') - q(k)q(k') \right)}{\sum_k k^2 q(k) - \left(\sum_k kq(k) \right)^2} \in [-1,1].$$

Subgraphs

Subgraphs

A convenient concept for us to look at now is that of a subgraph.

Definition:

A **subgraph** $G' = (V', E')$ of the graph $G = (V, E)$ is a graph such that $V' \subset V$ and $E' \subset E$.

Some further subgraph definitions

Some things to note:

- A small connected subgraph is also called a **motif**.
  - ⋆ The simplest non-trivial examples in undirected graphs are connected triples and triangles.
- A fully connected (complete) subgraph is called a **clique**.
  - ⋆ A clique is called maximal if we can't add another node to it so that it still remains a clique.
- A **spanning tree** is a tree subgraph that contains all vertices of the graph.
- A subgraph G' is called a **community**, if (for example; there are other definitions)

$$\sum_{i,j \in V'} a_{ij} > \sum_{i \in V', j \notin V'} a_{ij}.$$

  - ⋆ What this means is that the nodes in a community are more connected to each other than to the nodes outside it.

Some examples

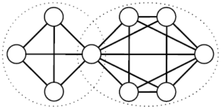

FIG. 20. Schematic of overlapping communities. Two communities are shown by dotted circles. One node belongs to both communities.

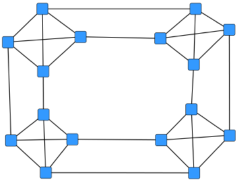

FIG. 21. Ring of 4 cliques.

Another figure from Lambiotte’s lecture notes

Lots of directions to go from here...

- **Motifs:** one of the main motivations to study clustering (we will do this next), counting motifs (cycles in particular), exponential random graphs.

- Motif analysis is particularly popular in the study of biological systems, where over-represented motifs are understood as building blocks of the network, that can be combined to form more complex structures.

- **Spanning trees** are related to Kirchhoff's matrix-tree theorem (related to eigenvalues of the *graph Laplacian*, which we will also see next)

- **Communities** are obviously linked to community detection, one of the most studied subjects in networks (if we have time, I will see if we can cover a bit of this)

Clustering

Clustering

Clustering aims to quantify the likelihood that two neighbours of a given vertex are themselves neighbours. There are two different definitions in the literature; the first one is easier for analytic computations, the second one for numerics.

Definition: global clustering coefficient

The **global clustering coefficient** for an undirected graph is defined as

$$C = \frac{3 \times \text{\# of (connected) triangles}}{\text{\# of (connected) triples}} = \frac{3 \sum_{i<j<l} a_{ij} a_{jl} a_{li}}{\sum_{i<j<l} (a_{ij} a_{il} + a_{ji} a_{jl} + a_{li} a_{lj})} \in [0, 1].$$

Definition: local clustering coefficient

Alternatively, we can define the **local clustering coefficient**

$$C_i = \frac{\text{\# of triangles containing vertex } i}{\text{\# of triples centered on vertex } i} = \frac{\sum_{j<l} a_{ij} a_{jl} a_{li}}{\sum_{j<l} a_{ij} a_{il}} \in [0, 1],$$

and use the average $\langle C_i \rangle = \frac{1}{N} \sum_i C_i$ to quantify clustering.

Some things to note

- If any pair of the neighbours of a vertex i is adjacent to each other so that they form a triangle, then $C_i = 1$.
    - ⋆ e.g. a complete graph has $C = \langle C_i \rangle = 1$.
- Similarly, if no pair of neighbours of vertex i is adjacent to each other, the neighbourhood of i is starlike, and $C_i = 0$.
    - ⋆ e.g. a tree has $C = \langle C_i \rangle = 0$.
- Higher-order clustering coefficients can be defined similarly, using different subgraphs as basis.

Graph Spectra

Graph spectra

Recall that the adjacency matrix of a graph G is a matrix A such that $a_{ij} = 1$ if there is an edge between i and j.

Definition

The **spectral density** of a graph $G = (V, E)$ is

$$\rho(\lambda) := \frac{1}{N} \sum_{i \in V} \delta(\lambda - \lambda_i) \quad \text{where} \quad \lambda_1, \ldots, \lambda_N \in \mathbb{C}$$

are the eigenvalues of the adjacency matrix A.

Recall the **Perron-Frobenius theorem** (seen for Markov chains): A has a real eigenvalue $\lambda_1 > 0$ with maximal absolute value, and with corresponding real, non-negative eigenvector(s).

If the graph is connected, this eigenvalue has multiplicity 1 and $|\lambda_j| < \lambda_1$ for all other eigenvalues with $j \neq 1$.

More about adjacency matrices and their eigenvalues

Knowing the spectral properties of adjacency matrices can be quite useful! For example:

- For an undirected graph, $(A^n)_{ij}$ is equal to the **number of walks** (paths which allow repeated vertices) from i to j of length n.
- We also have

$$\mathrm{Tr}(A^n) = \sum_{i=1}^{N} \lambda_i^n \qquad \text{and} \qquad \big(\mathrm{Tr}(A)\big)^n = 0,$$

  which can be used to derive statements like:

  ⋆ $$\sum_{i<j} \lambda_i \lambda_j = -|E|,$$

  ⋆ $$\sum_{i<j<l} \lambda_i \lambda_j \lambda_l = 2 \cdot \# \text{ of triangles in } G.$$

Other measures on graphs - centrality measures

The goal of centrality measures is to **quantify the importance of nodes in a network**.
For example,

- In social networks, one can use a combination of centrality measures to predict the future impact of a new user.
- In epidemiology, nodes with a high centrality are often targeted, e.g. by means of vaccination, in order to slow down the progress of a disease.
- In marketing, central users can be targeted to seed viral campaigns.

We already saw an example: the degree, or **degree centrality** (nodes with higher degrees are considered to be important). The degree is effective in various situations but not always, and this has motivated the introduction of different types of centrality measures.

Distance based centrality measures

The first (and probably most popular) two examples of centrality measures are based on the distance between pairs of nodes:

- The **closeness centrality for node *i*** (only defined for connected graphs) is the inverse of the mean distance from node *i* to any other node and is denotned by

$$\text{closeness}_i = \frac{N-1}{\sum_{j=1,j\neq i}^{N} d_{ij}}.$$

- The **betweenness centrality** of node *i* is the fraction of the shortest paths in the graph passing through *i*.

$$\text{betweenness}_i = \frac{2}{(N-1)(N-2)} \sum_{j=1,j\neq i}^{N} \sum_{\ell=1,\ell\neq i}^{j-1} \frac{\sigma_{j\ell}^{i}}{\sigma_{j\ell}}.$$

Here $\sigma_{j\ell}$ is the number of the shortest paths between j and ℓ, and $\sigma_{j\ell}^{i}$ is the number of such paths that pass through i.

A path based centrality measure

In some cases (e.g., communication and infectious diseases), short walks can be more important than long walks.
We know that given an adjacency matrix A of the graph G, the number of walks from i to j with ℓ steps is given by the $(i, j)-$th element of A^ℓ.

So, to measure the “importance" of a node i, we can see how many walks there are from i to j and scale the weight of each walk of length $\ell \geq 0$ by a factor of α^ℓ, where $0 < \alpha < 1$.
With this, we define a weighted sum of the number of walks from i to j, the (i, j) element of

$$I + A + \alpha^2 A^2 + \cdots = (I - \alpha A)^{-1}.$$

The **Katz centrality** of the $i-$th node is then defined by

$$\mathrm{Katz}_i = \sum_{j=1}^{N} \left[(I - \alpha A)^{-1} \right]_{ij}$$

Note that: if $\alpha = 0$, then $\mathrm{Katz}_i = 1$ for all i. Therefore, we are interested in making α large to diversify the values of Katz_i.

PageRank

Our final example is one you might be familiar with.

The **PageRank** is a well-known centrality measure for directed networks.

- It was first introduced for ranking webpages and later adopted in a variety of applications. (e.g. it is the algorithmo used by Google to rank webpages showing on your searches!)
- It is defined as the stationary density of a discrete-time random walk on directed networks.
- It is defined by a typical **recursive metric**, based on the circular idea that a node is important if it receives connections from many important nodes. This relation leads to an eigenvector problem.
- Similar arguments lead to other centrality measures, such as Eigenvector centrality.

If we have time, we can have a closer look at this on week 10!

The Graph Laplacian

Graph Laplacian

Definition

The **Graph Laplacian** for a graph $G = (V, E)$ with adjacency matrix A is defined as

$$Q := A - D \quad \text{where} \quad D = \left(\delta_{ij} \sum_{l \neq i} a_{il} : i, j \in V\right).$$

Note that:

- You can think of this as if Q defines a generator matrix of a continuous-time random walk on V with transition rates a_{ij}.
- If G is a weighted graph, we can define Q in a similar way, with A replaced by $W = (w_{ij})$.
- So we can see that using weighted graphs, any finite state CTMC can be represented in this way.
- The Laplacian determines the first order linearised dynamics of many complex processes on graphs and is therefore of particular importance.

Some properties of the graph Laplacian

The Graph Laplacian tells us a lot about networks. For example:

- Q has eigenvalues in $\mathbb{C}$ with real part $\mathrm{Re}(\lambda) < 0$ except for $\lambda_1 = 0$.
  - ⋆ This follows directly from the Gershgorin theorem, when you notice the vanishing row sums.
  - ⋆ For undirected graphs, A (and Q) is symmetric, therefore all the λ_i's are real.
- When looking at Graph Laplacians, it is customary to order its eigenvalues from the largest to the smallest real part.
  - ⋆ The **multiplicity of λ_1 gives us the number of connected components** in undirected graphs.
  - ⋆ Properly chosen orthogonal eigenvectors to λ_1 have non-zero entries on the individual connected components.
- The “second" eigenvalue of Q is often called the **spectral gap**. It determines the relaxation time of the dynamics induced by Q.
  - ⋆ The corresponding eigenvector, u_2, is called the **Fiedler vector**. The smaller the second largest real part of an eigenvalue, the harder it is to cut G into separated components by removing edges.

Some comments about spectral properties of graphs

Spectral properties of networks can be used to reveal important structural properties of these graphs, and are at the core of several algorithms for community detection or clustering.

- For degree centrality, the dominant eigenvector of the normalised Laplacian is essentially equivalent to PageRank.
    - ⋆ The normalised Laplacian has its entries normalised by the degrees of corresponding nodes
- Likewise, the dominant eigenvector of the adjacency matrix is called eigenvector centrality.
- The Fiedler vector is also associated to important structural patterns, like the presence of botllenecks and communities in the network.