

Lecture 1

Graph Theory

1.1 Adjacency Matrix

As we saw earlier, many systems of interest can be seen as being constituted by several components interacting among themselves. We call each element a “**node**” and each connection an “**edge**”. When the structure of the nodes is less important than the connections between them, we say we are in the presence of a “Graph” or “Network”, structures that are studied by the branch of mathematics known as Graph Theory.

Mathematically, a graph can be represented by a matrix $A \equiv (a_{ij})$ where matrix element a_{ij} is meant to representing a link from i to j . The so called “Adjacency Matrix”:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} \end{bmatrix}$$

where:

$$a_{ij} = \begin{cases} 1 & ij \text{ connected} \\ 0 & ij \text{ unconnected} \end{cases}$$

provides us with a compact way of representing all the connectivity information about the system. See Figure 1.1 for a simple example. In the remainder of this lecture we’ll use this matrix to derive several other useful properties of networks.

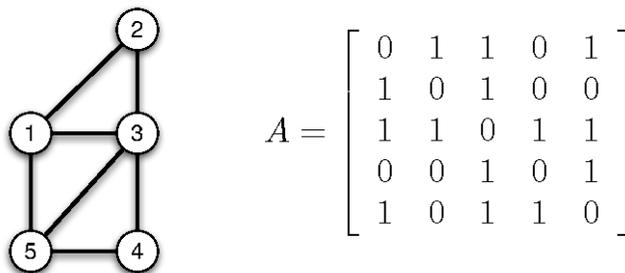


Figure 1.1: A simple network and the corresponding adjacency matrix.

1.2 Degree Distribution

When studying real world networks, one is usually more interested in statistical distributions than in the detailed structure of how each component connects with every other one. Of particular interest is the degree distribution of the nodes. The degree of a node i , k_i , is simply the number of connections node i has. This quantity can be easily calculated from the adjacency matrix:

$$k_i = \sum_j a_{ij}$$

A vector containing the degrees of each node is then given by:

$$\vec{k} = A \cdot \vec{1}$$

where $\vec{1}$ is the vector where every element is identically one. The degree distribution is then defined as:

$$P(k) = \frac{1}{N} \sum_i \delta(k_i - k)$$

where N is the total number of nodes, and is simply the probability that a randomly chosen node has degree k .

1.3 Paths and Cycles

From our definition, we know that when $a_{ij} = 1$ there must be a direct link connecting nodes i and j . Similarly, if $a_{jk} = 1$ then there is a link connecting j and k , in which case there is a way to go from i to k in two steps, by going through node j . Mathematically, we can represent a length two path by an element:

$$p_{ik} = a_{ij}a_{jk}$$

This definition provides us with a simple way of calculating how many ways there are of going from i to k in two steps:

$$p_{ik} = \sum_j a_{ij}a_{jk}$$

which we quickly recognize as being the formula to calculate element $p_{ij}^{(2)}$ of matrix:

$$P^{(2)} = A \cdot A \equiv A^2$$

with $P^{(2)}$ standing for “Path of length 2”. For the small matrix in our example, we obtain:

$$A^2 = \begin{bmatrix} 3 & 1 & 2 & 2 & 1 \\ 1 & 2 & 1 & 1 & 2 \\ 2 & 1 & 4 & 1 & 2 \\ 2 & 1 & 1 & 2 & 1 \\ 1 & 2 & 2 & 1 & 3 \end{bmatrix}$$

which tells us that there are 2 ways of going from node 2 to node 5 in two steps (once via node 3 and once via node 1). One thing to notice about this matrix, is the fact that the diagonal elements

$a_{ii}^{(2)} \neq 0$. According to our definition, these elements represent the number of ways one can go from node i not node i in two steps, or, in other words, the degree of node i . We convince ourselves of this by considering that we can follow each link connected to i to a neighbour and then follow it back again to return to i .

A similar train of thought can be extended to paths of arbitrary length. A path going from node i to node j in n steps is then given by:

$$a_{ik}a_{kl} \cdots a_{lj}$$

and the total number of paths is:

$$\sum_{k,l} a_{ik}a_{kl} \cdots a_{lj} = (A^n)_{ij}$$

In general, any path of length n that has the same node as both start and end points is called a “Cycle of length n ”.

1.4 Clustering coefficient

The degree of a node gives us information about how well connected a node is. However, in many situations, we are also interested in knowing how dense the connections between the neighbors of a node are. One way of measuring this quantity is to count how many connections there are between the k_i neighbors of node i out of all possible ones:

$$C_i = \frac{\# \text{ links}}{\text{all possible links}}$$

C_i , known as the clustering coefficient of node i , varies between 0 and 1. 0 when there are no connections between its neighbors and 1 when all connections are present. The total number of possible pairwise connections between k_i nodes is simply:

$$\binom{k_i}{2} \equiv \frac{k_i(k_i - 1)}{2}$$

but how can we measure the number of connections that are present? For this let us remember the number of cycles of length 3:

$$p_{ii}^{(3)} = \sum_{jk} a_{ij}a_{jk}a_{ki}$$

since by definition, j and k are neighbors of i this formula will only have non-zero terms when $a_{jk} \equiv 1$. In other words, the number of cycles of length 3 gives us **twice** the number of connections between the neighbors of i . The clustering coefficient is then:

$$C_i = \frac{p_{ii}^{(3)}}{k_i(k_i - 1)}$$

If we take the average of this quantity over all the nodes in the network, we obtain the average clustering coefficient,:

$$\langle C \rangle = \frac{1}{N} \sum_i C_i$$

that quantifies how densely connected a given network is.

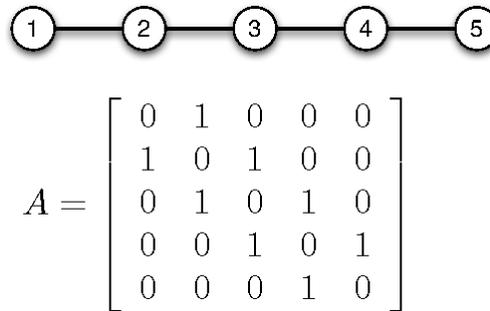


Figure 1.2: Linear network and its adjacency matrix.

1.5 Network Diameter

The paths defined above give us the number of ways there are to go from node i to node j in **exactly** n steps. Using this definition, we can also measure what is the **shortest path** that connects two nodes: the first value of n for which $p_{ij}^{(n)} \neq 0$. As an example, consider the network plotted in Fig. 1.2.

It's clear that the length of the shortest path between nodes 1 and 5 is four, since we have to cross 3 nodes and 4 edges to reach from one to the other. In fact, we can easily see that the corresponding element of A^4 is non-zero:

$$A^4 = \begin{bmatrix} 2 & 0 & 3 & 0 & 1 \\ 0 & 5 & 0 & 4 & 0 \\ 3 & 0 & 6 & 0 & 3 \\ 0 & 4 & 0 & 5 & 0 \\ 1 & 0 & 3 & 0 & 2 \end{bmatrix}$$

also note that several elements of this matrix are zero. For example, $(A^4)_{14} \equiv 0$ since there is no way to go from node 1 to node 4 in exactly four steps (only 3 or 5 if you allow for one instance of backtracking).

The longest shortest path that is present in a network is known as the **Network Diameter**. Even though there is no simply way to calculate mathematically this quantity given an adjacency matrix, there are several numerical algorithms that can achieve this task. One of the most common ones is known as Dijkstra's algorithm, after its creator Dutch computer scientist Edsger Dijkstra.

1.6 Small World

As an application of the concepts introduced so far, let us consider the case of the Watts-Strogatz network introduced by Duncan Watts and Steven Strogatz in 1998. The network they propose is deceptively simple. A simply circle where each node is connected with K of its nearest neighbors.

A network this simple shouldn't present any particularly surprising properties. However, the authors took their model a step further and **randomly** rewired each node in the network with probability p and studied how the properties of this random network changed with p .

In Fig. 1.3 we plot several configurations of the network as a function of p . It should be clear that for $p = 0$ nothing changes in the network. The network remains completely regular, with

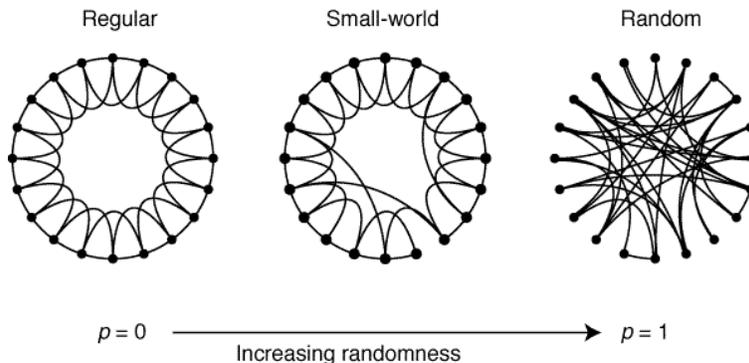


Figure 1.3: Watts Strogatz network. After Nature 393, 440 (1998)

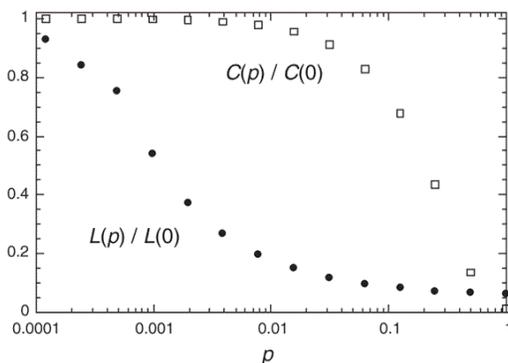


Figure 1.4: Small world effect. After Nature 393, 440 (1998)

a relatively large clustering coefficient and large diameter. In the opposite extreme, $p = 1$, the network becomes completely **random**. Since any two nodes can be connected regardless of their neighbors, the clustering coefficient becomes particularly small as does the diameter.

Despite the simplicity of these two limits, the intermediate regime is particularly interesting. As soon as one rewires a few edges (small p), the diameter quickly decreases since we are, in effect, introducing shortcuts in the network. On the other hand, the clustering coefficient is relatively insensitive to these added shortcuts and remains large until p reaches significantly large values. The behavior of these two quantities as a function of p is plotted in Fig. 1.4.

The intermediate regime is known as “Small World” and represents a simple model to explain an observation made by American Sociologist, Stanley Milgram, in 1967. Using a simple experiment, Milgram was able to measure the diameter of the social network connecting americans and found it to be only 6 hops long. An observation that came to be known as **six degrees of separation**.