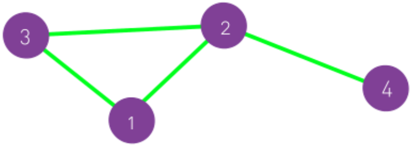
Advanced Algorithms and Computational Models (module A)

Elements of graph theory

Giacomo Fiumara
gfiumara@unime.it

A key property of each node is its *degree*, representing the number of links it has to other nodes
k_i is the degree of the i^{th} node in the network

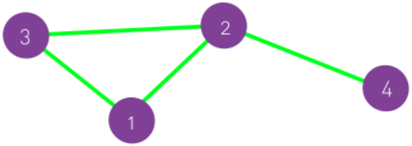


In this case, we have $k_1 = 2$, $k_2 = 3$, $k_3 = 2$, $k_4 = 1$

In an undirected network, the *total number of links L* can be expressed as the sum of the node degrees:

$$L = \frac{1}{2}\sum_{i=1}^{N} k_i$$

Here the factor $\frac{1}{2}$ corrects for the fact that, in the sum, each link is counted twice



Average (mean)

$$\langle x \rangle = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

n^{th} **moment**

$$\langle x^n \rangle = \frac{x_1^n + x_2^n + \cdots + x_N^n}{N} = \frac{1}{N} \sum_{i=1}^{N} x_i^n$$

Standard deviation

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}$$

Distribution of x

$$p_x = \frac{1}{N}\delta_{x,x_i}$$

where

$$\sum_i p_x = 1 \; or \; \int p_x dx = 1$$

For undirected networks, average degree is defined as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i = \frac{2L}{N}$$

In directed networks we distinguish between *incoming degree* k_i^{in}, representing the number of links that point to node i, and *outgoing degree* k_i^{out}, representing the number of links that point from node i to other nodes

A node's *total degree* k_i is given by

$$k_i = k_i^{in} + k_i^{out}$$

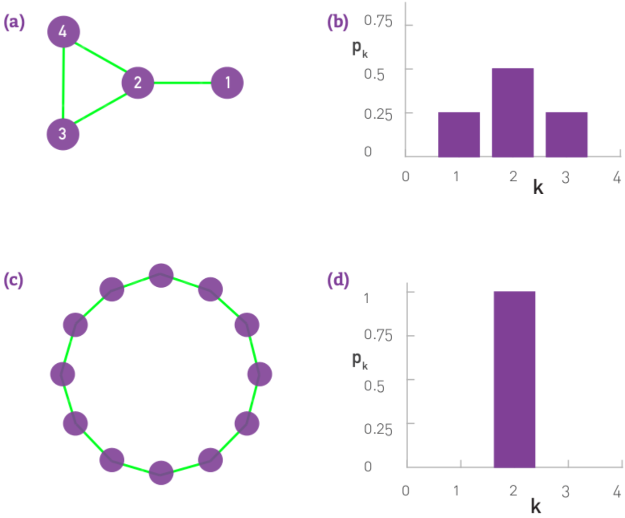
The *degree distribution* p_k provides the probability that a randomly selected node in the network has degree k. Since p_k is a probability, it must be normalized, i.e.

$$\sum_{i=1}^{\infty} p_k = 1$$

For a network with N nodes, the degree distribution is the normalized histogram given by

$$p_k = \frac{N_k}{N}$$

where N_k is the number of degree-k nodes.



(a)
4
2
1
3
(b)
0.75
p_k
0.5
0.25
0
0
1
2
k
3
4
(c)
(d)
1
0.75
p_k
0.5
0.25
0
0
1
2
k
3
4

– The degree distribution has assumed a central role in network theory following the discovery of scale-free networks.

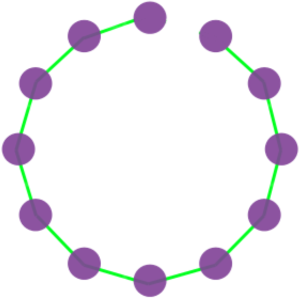
– The calculation of most network properties requires to known p_k

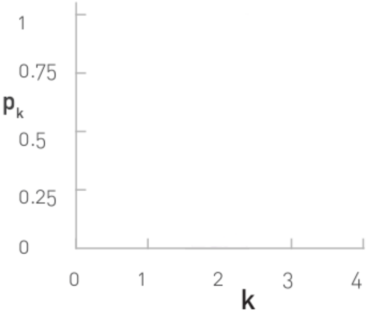
– For example, the average degree of a network can be written as:

$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

– Another reason is that the precise functional form of p_k determines many network phenomena, from network robustness to the spread of viruses

A simple exercise:



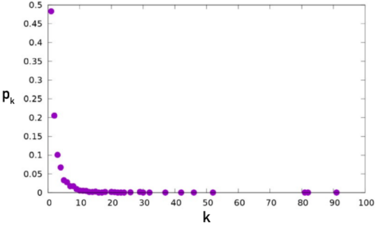


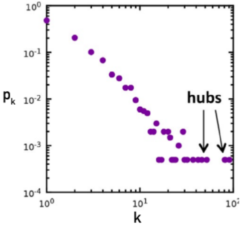
Degree, Average Degree and Degree Distribution

Degree Distribution

Degree, Average Degree and Degree Distribution

Degree Distribution





Adjacency Matrix

– A complete description of a network requires to keep track of its links

– The simplest way to achieve this is to provide a complete list of the links

– For mathematical purposes it is better to represent a network through its adjacency matrix

– The *adjacency matrix* of a directed network of N nodes has N rows and N columns

– For directed networks the sums over the rows and columns of the adjacency matrix provide the incoming and outgoing degrees, that is

$$2L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} = \sum_{ij}^{N} A_{ij}$$

Adjacency Matrix

– The number of nonzero elements of the adjacency matrix is $2L$, or twice the number of links
– Indeed, an undirected link connecting two nodes i and j appears in two entries:

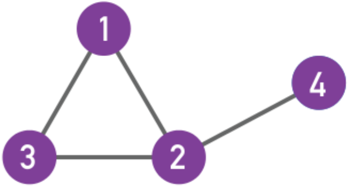
$$A_{ij} = A_{ji} = 1$$

Adjacency Matrix

$$A_{ij} = \begin{matrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{matrix}$$

Adjacency Matrix

Undirected Network



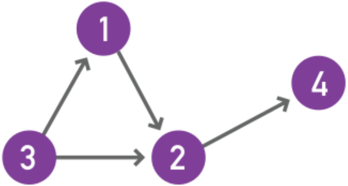
$$A_{ij} = \begin{matrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

$$k_2 = \sum_{j=1}^{4} A_{2j} = \sum_{i=1}^{4} A_{i2} = 3$$

$$A_{ij} = A_{ji} \;\; A_{ii} = 0$$

$$L = \frac{1}{2} \sum_{i,j=1}^{N} A_{ij}$$

$$\langle k \rangle = \frac{2L}{N}$$



$$k_2^{in} = \sum_{j=1}^{4} A_{2j} = 2 \;\; k_2^{out} = \sum_{i=1}^{4} A_{i2} = 1$$

$$A_{ij} \neq A_{ji} \;\; A_{ii} = 0$$

$$L = \sum_{i,j=1}^{N} A_{ij}$$

$$\langle k^{in} \rangle = \langle k^{out} \rangle = \frac{L}{N}$$

$$A_{ij} = \begin{matrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{matrix}$$

Real Networks are Sparse

– In real networks, the number N of nodes and L of links can vary widely

– For example, the neural network of the worm *Caenorhabditis Elegans*, whose nervous system has been completely mapped, has $N = 302$ neurons (nodes)

– In contrast, the human brain is estimated to have about $N = 10^{11}$ neurons

– The genetic network of a human cell has about 20000 genes (nodes)

– The social network consists of $N = 7 \times 10^9$ individuals

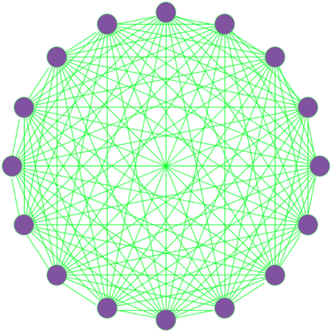
– The WWW is estimated to have $N > 10^{12}$ documents

Real Networks are Sparse

– In a network of N nodes, the number of links can change between $L = 0$ and L_{max}, where

$$L_{max} = \frac{N(N-1)}{2}$$

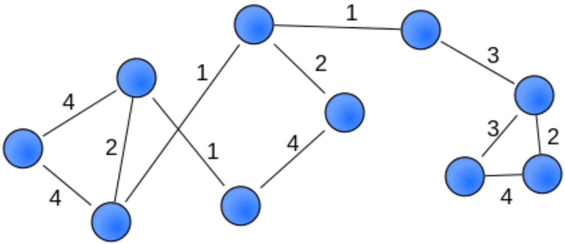
is the total number of links present in a *complete graph* of size N



Weighted Networks

– In many applications it is necessary to study *weighted networks*, where each link (i, j) has a unique weight w_{ij}

– For example, in mobile call networks the weight can represent the total number of minutes two individuals talk with each other on the phone, or the total number of calls

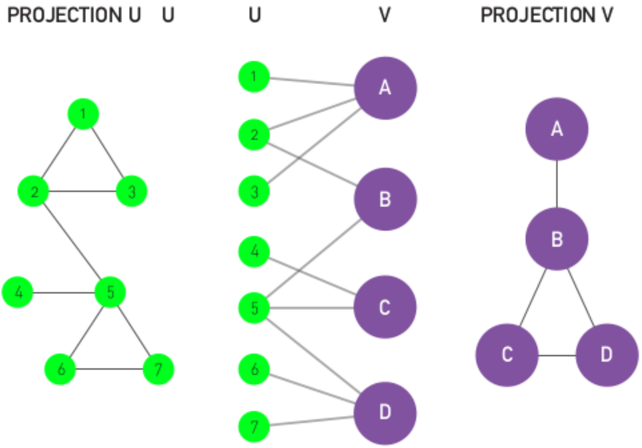


Bipartite Networks

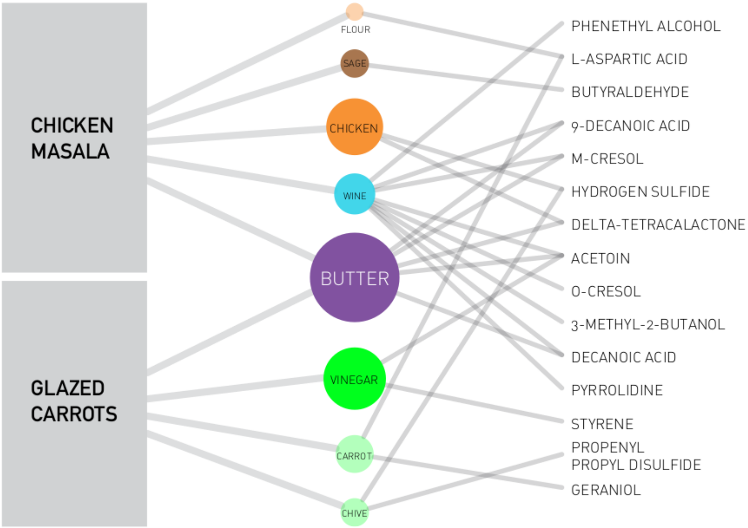
– A *bipartite network* is a network whose nodes can be divided into two disjoint sets U and V such that each link connects a U-node to a V-node

– Two *projections* can be generated for each bipartite network. The first connects two U-nodes if they are linked to the same V-node in the bipartite representation. The second projection connects the V-nodes by a link if the connect to the same U-node

Bipartite Networks



Bipartite Networks

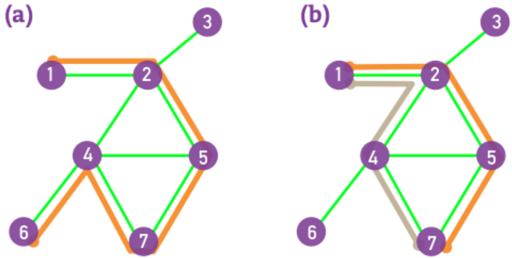


Paths and Distances

– In networks, the concept of physical distance is meaningless

– For example, two web pages may reside on computers physically very distant but have a link to each other

– In networks, physical distance is replaced by *path length*. A *path* is a route that runs along the links of the network. The path length is the number of links the path contains

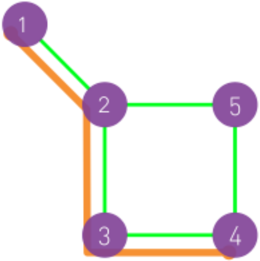


Paths and Distances

Shortest Path

- The shortest path between nodes i and j is the path with the fewest number of links
- The shortest path is often called the *distance* between nodes i and j and is denoted as d_{ij}
- We can have multiple shortest paths of the same length between a pair of nodes
- In an undirected network, $d_{ij} = d_{ji}$, i.e. the distance between node i and i is the same as the distance betweem node j and i
- In a directed network, the existence of a path from node i to node j does not guarantee the existence of a path from j to i

- The *diameter* of a network, denoted by d_{max}, is the maximum shortest path in the network. It is the largest distance recorded between any pair of nodes

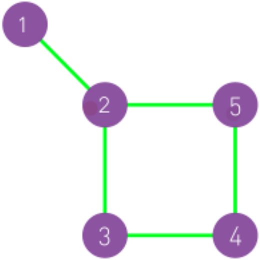


$d_{1\to 4}=3=d_{max}$

Paths and Distances

Shortest Path

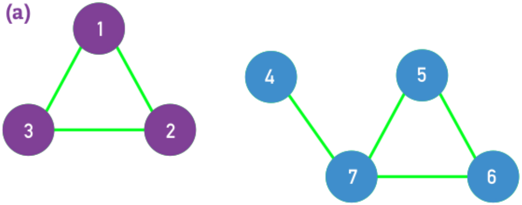
– The *average path length*, denoted by $\langle d \rangle$ is the average distance between all pairs of nodes in the network.



$$\langle d \rangle = [d_{1\to2} + d_{1\to3} + d_{1\to4} + d_{1\to5} + \\ + d_{2\to3} + d_{2\to4} + d_{2\to5} + \\ + d_{3\to4} + d_{3\to5} + \\ + d_{4\to5}]/10 = 1.6$$

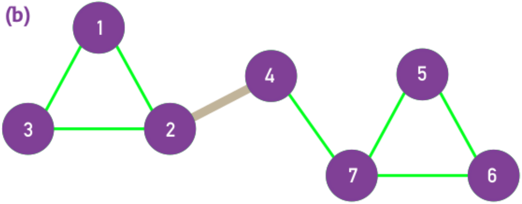
Connectedness

- Directly related to path and distances is the concept of *connectedness*
- In an undirected network, nodes i and j are *connected* if there is a path between them. They are *disconnected* if such a path does not exist, in which case we have $d_{ij} = \infty$
- A network is connected if **all** pairs of nodes in the network are connected
- A network is disconnected if there is **at least** one pair with $d_{ij} = \infty$
- If a network is disconnected, its subnetworks are called *components* or *clusters*



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

- If a network consists of two components, a properly placed link can connect them, making the network connected. Such a link is called *bridge*



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

Clustering Coefficient

- The clustering coefficient captures the degree to which the neighbours of a given node link to each other
- For a node i with degree k_i, the local clustering coefficient is defined as

$$C_i = \frac{2L_i}{k_i(k_i - 1)}$$

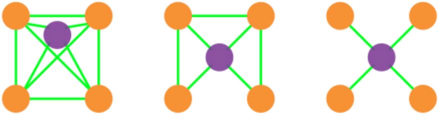
where L_i represents the number of links between the k_i neighbours of node i.

Clustering Coefficient

C_i is the probability that two neighbourse of a node link to each other. It measures the network's local density

C_i is between 0 and 1:

- If $C_i = 0$, none of the neighbours of node i link to each other
- If $C_i = 1$, the neighbours of node i form a complete graph, they all link to each other

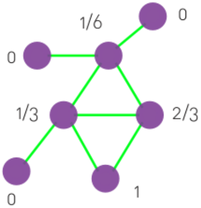


C_i=1 C_i=1/2 C_i=0

Clustering Coefficient

– The degree of clustering of a whole network is captured by the *average clustering coefficient*, $\langle C \rangle$, representing the average of C_i over all nodes $i = 1, \ldots, N$

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} C_i$$



$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

$$C_{\triangle} = \frac{3}{8} = 0.375$$