NETWORKS AND GRAPHS

Data and Networks
Part II

Dr. Franck Kalala Mutombo

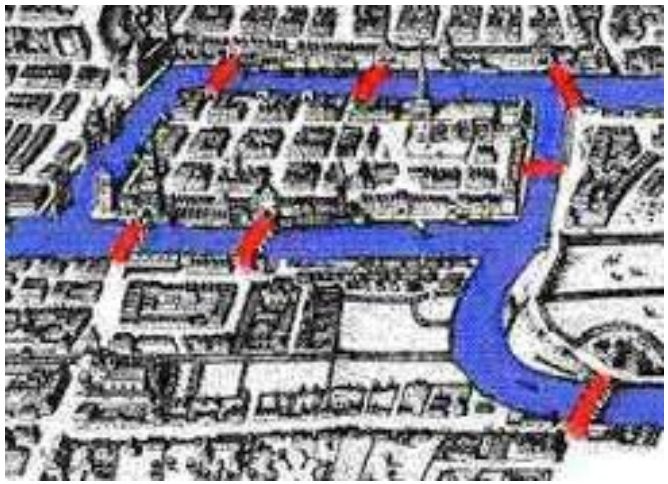
African Institute for Mathematical Sciences (AIMS) South Africa

franckm@aims.ac.za

May 16, 2017

Graph theory: The Bridges of Konigsberg

Konigsberg, the capital of Eastern Prussia (Russia) in 1735.

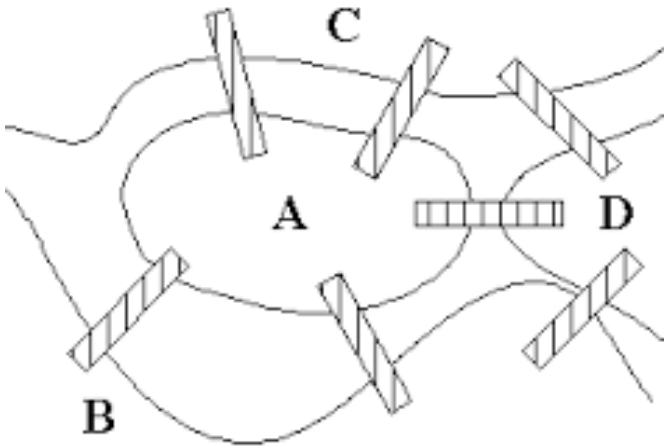


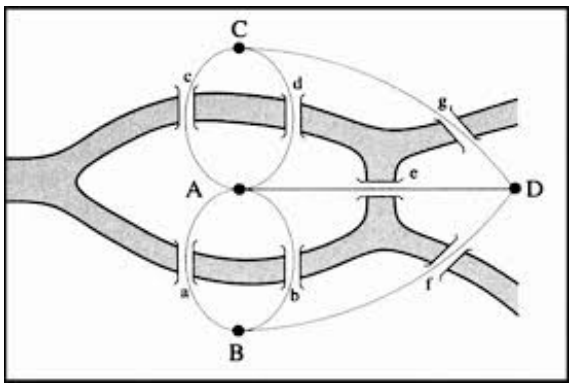
puzzle

Can one walk across all seven bridges and never cross the same one twice? Despite many attempts, no one could find such path.

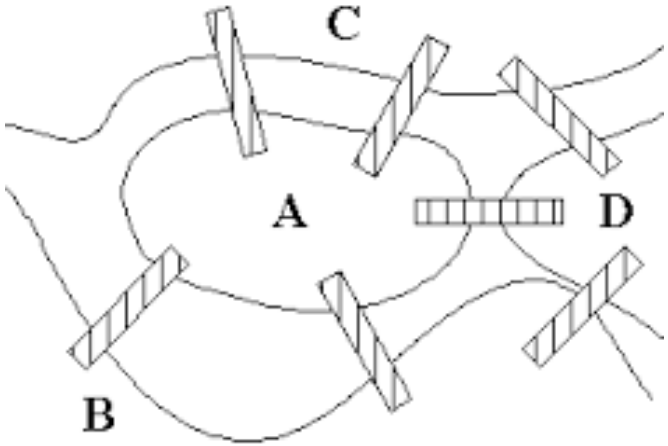
puzzle

The problem remained unsolved until 1735, when Leonard Euler (12-13) gave a rigorous mathematical proof showing that such a path does not exist.



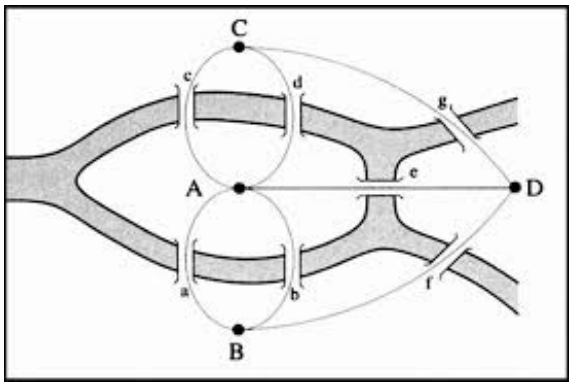


1. Euler represented each of the four lands area separated by the river by letter A,B,C and D.
2. Next he connected with lines each piece of land that had a bridge between them.



Euler observation

If there is a path crossing all bridges, but never the same bridge twice, then nodes with odd number of links must be either the starting of the end point of this path.



♣

In fact, if one arrive at a node with an odd number of links, he may find himself having no unused link for him to leave it.

Can you solve the puzzle?

Euler showed that there is no continuous path that would cross the seven bridges while never crossing the same bridge desired path.

Graph theory: The Bridges of Konigsberg

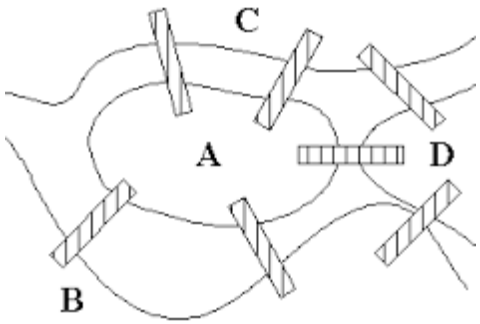
Euler has the merit to solve for the first time a mathematical problem using a graph. We learn two important things from Euler solution:

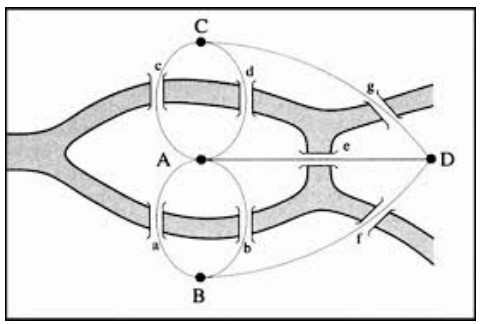
1. Some problems become simple and easy to deal with when then are represented as a graph.
2. The existence of the path does not depend on our ingenuity to find it. Rather, it is a property of the graph.

Networks have properties encoded in their structure that limit or enhance their behaviour.

Graph theory

To understand how networks can affect properties of system, one need to be familiar with graph theory. In this chapter we will be learning the following





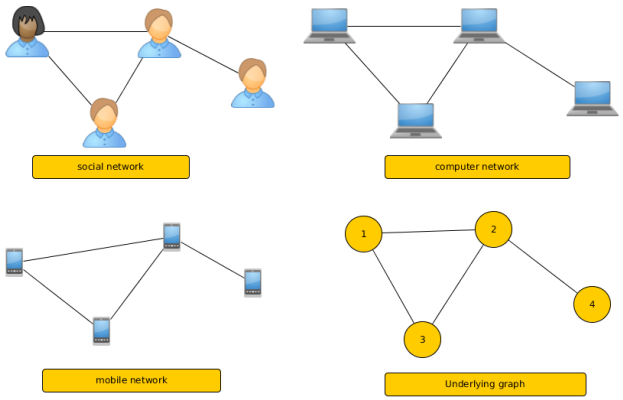
1. how to represent a networks as a graph,
2. introduce some network characteristics,
3. degrees, degree distribution,
4. clustering,
5. average path length,
6. weigthed networks,
7. directed networks,
8. bipartite networks, etc.

Networks and Graphs

1. If we want to understand a complex system, we first need to know how its components interact with each other.
2. In other words we need a map of its wiring diagram.
3. A network is a catalog of a system's components often called nodes or vertices and the direct interactions between them, called links or edges.
4. This network representation offers a common language to study systems that may differ greatly in nature, appearance, or scope.

Networks and Graphs

Different networks same graphs. Three different systems have exactly the same network representation.



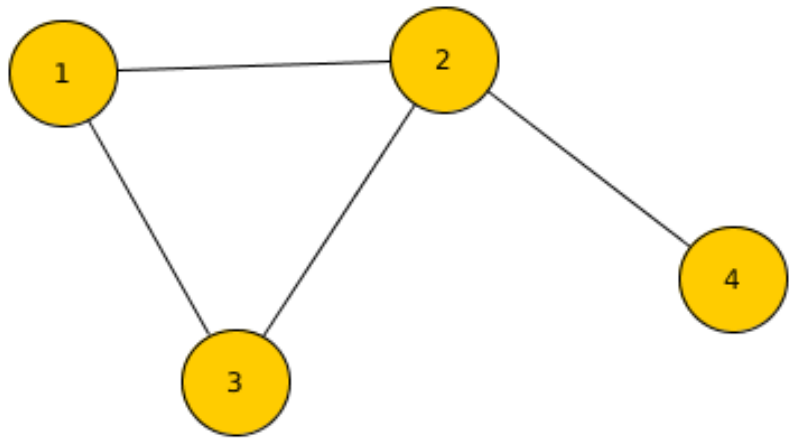
1. social network where nodes are people and links are friendship, relationship,
2. computer network where nodes are computers and links are protocols,
3. mobile network where nodes are cell phones and links are

While the nature of the nodes and the links differs, these networks have the same graph representation, consisting of $N = 4$ nodes and $L = 4$ links.

Networks and Graphs: basic characteristics

1. Number of nodes, or N, represents the number of components in the system. We will often call N the size of the network. To distinguish the nodes, we label them with $i = 1, 2, \cdots, N$.
2. Number of links, which we denote with L, represents the total number of interactions between the nodes.

Links are rarely labeled, as they can be identified through the nodes they connect. For example, the (2, 4) link connects nodes 2 and 4.



$N = 4$ and $L = 4$

Networks or Graphs?

In the scientific literature the terms network and graph are used interchangeably:

Network Science	Graph Theory
Network	Graph
Node	Vertex
Network	Graph

Networks or Graphs?

The terminology $\{network, node, link\}$ often refers to real systems such

1. The WWW is a network of web documents linked by URLs;
2. Society is a network of individuals linked by family, friendship or professional ties;
3. The metabolic network is the sum of all chemical reactions that take place in a cell.

and the terminology $\{graph, vertex, edge\}$ is used when discussing mathematical representation of these networks:

1. web graph
2. social graph (Facebook)
3. metabolic graph

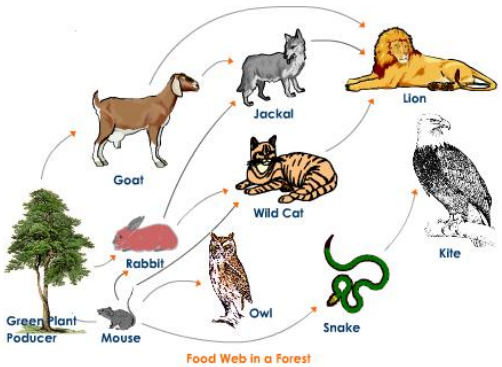
this distinction is rarely made, so these two terminologies are often synonyms of each other.

Directed, undirected networks

Definition

A network is called directed (or digraph) if all of its links are directed; it is called undirected if all of its links are undirected. Some networks simultaneously have directed and undirected links.

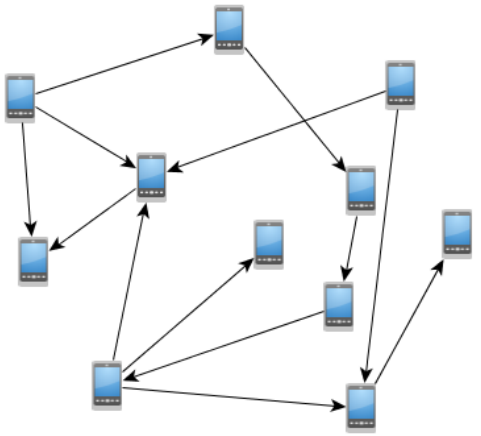
The links of a directed networks have direction while the links in a undirected networks have no directions.



For example in the foodweb network links are directed and in a social network links are undirected.

Directed, undirected networks

Some biological reactions are reversible (i.e., bidirectional or undirected) and others are irreversible, taking place in only one direction (directed).



Some systems have directed links, like in the cell network where links are phone calls, in which one person calls the other.
Other systems have undirected links, like romantic ties: if I date Janet, Janet also dates me, or like transmission lines on the power grid, on which the electric current can flow in both directions.

Directed, undirected networks

The choices we make when we represent a system as a network will determine our ability to use network science successfully to solve a particular problem.

The way we define the links between two individuals dictates the nature of the questions we can explore:

Example

1. By connecting individuals that regularly interact with each other in the context of their work, we obtain the organizational or professional network, that plays a key role in the success of a company or an institution, and is of major interest to organizational research.

Directed, undirected networks

Example

1. By linking friends to each other, we obtain the friendship network, that plays an important role in the spread of ideas, products and habits and is of major interest to sociology, marketing and health sciences.
2. By using phone and email records to connect individuals that all or email each other, we obtain the acquaintance network, capturing a mixture of professional, friendship or intimate links, of importance to communications and marketing.
3. By connecting individuals that have an intimate relationship, we obtain the sexual network, of key importance for the spread of sexually transmitted diseases, like AIDS, and of major interest for epidemiology.

Network theory and system

When modelling a system using network theory, careful considerations must precede our choice of nodes and links, ensuring their significance to the problem we wish to explore.

characteristics of common data sets used in network science by researchers

Network	Nodes	Links	Directed / Undirected	N	L	‹K›
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.34
WWW	Webpages	Links	Directed	325,729	1,497,134	4.60
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594	2.67
Mobile-Phone Calls	Subscribers	Calls	Directed	36,595	91,826	2.51
Email	Email addresses	Emails	Directed	57,194	103,731	1.81
Science Collaboration	Scientists	Co-authorships	Undirected	23,133	93,437	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Papers	Citations	Directed	449,673	4,689,479	10.43
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802	5.58
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930	2.90

1. social systems (mobile call graph or email network),
2. collaboration and affiliation networks (science collaboration network, Hollywood actor network)
3. information systems (WWW)
4. technological and infrastructural systems (Internet and power grid)
5. biological systems (protein interaction and metabolic network)
6. reference networks (citations)

Degree

Definition, [Neighborhood ($N_G(v)$)]

The neighborhood of a vertex $v \in V$ is a set of all vertices that are adjacent to v. Mathematically, $N_G(v) = \{u \in V | uv \in E\}$.

Degree

The degree of a vertex v is the number of edges incident to it. A self-edge is counted as two edges. The degree of a node v is the number of nearest neighbours of v, that is, $k_v = |N_G(v)|$.

The degree is one of the key property of a node. It can also be seen a the number of links a node has to other node.

example

For example in a cell network, the degree can represent the number of mobile phone contacts an individual has in the call graph (i.e. the number of different individuals the person has talked to), or the number of citations a research paper gets in the citation network.

Degree

Leaf

If $k_v = 0$, then node v is said to be isolated in G, and if $k_v = 1$, then v is a leaf of the graph.

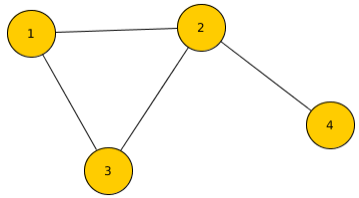
minimum degree

The minimum degree $k_{min}(G) = min\{k_v | v \in G\}$ and the maximum degree $k_{max}(G) = max\{k_v | v \in G\}$.

maximun degree

For a directed network, we consider two types of degrees, namely in-degree (k_v^{in}) and the out-degree (k_v^{out}), which are the number of edges pointing towards or departing from a node v respectively. The total degree k_v is $k_v = k_v^{in} + k_v^{out}$.

Degree



Nodes degrees

$$k_1 = 2\ , k_2 = 2\ , k_3 = 2\ , k_4 = 1.$$

Handshaking lemma

For any given undirected network $G = (V, E)$, where V is the set of nodes and E the set of edges, the sum of all vertex degrees is equal to twice the number of edges.

$$\sum_{v \in V} k_v = 2\ |E|. \tag{1}$$

Therefore, the total number of links, L, in an undirected network can be expressed in term of the sum of the node degrees:

$$L = |E| = \frac{1}{2} \sum_{v=1}^{N} k_v \tag{2}$$

Average degree

The average degree is an important property of a network and defined as

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{3}$$

The total number of links in a directed network is

$$L = \sum_{i=1}^{N} k_i^{in} = \sum_{i=1}^{N} k_i^{out} \tag{4}$$

The 1/2 factor is now absent, as for directed networks the two sums in (4) separately count the outgoing and the incoming degrees. The average degree of a directed network is

$$\langle k^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{in} = \langle k^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i^{out} \tag{5}$$

Descriptive statistic

For a sample having n units, $x_1, \cdots, x_N$ we have these four parameter that characterise it:

$$\text{Average (mean)} \quad \langle k \rangle = \frac{x_1 + \cdots + x_n}{N} = \frac{1}{N} \sum_i^N x_i \tag{6}$$

$$n^{th} \text{ moment} \quad \langle k^n \rangle = \frac{x_1^n + \cdots + x_n^n}{N} = \frac{1}{N} \sum_i^N x_i^n \tag{7}$$

$$\text{standard deviation} \quad \sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \langle x \rangle)^2} \tag{8}$$

$$\text{distribution de } x \quad p_x = \frac{1}{N} \sum_i \delta_{x,x_i} \tag{9}$$

$$\sum p_x = 1 \quad \left(\int p_x dx = 1 \right) \tag{10}$$

Degree distribution

.

The degree distribution of a network is obtained in terms of the probability p_k and is defines as the probability that a node chosen uniformly at random has degree k or equivalently as the fraction of nodes in the graph having degree k.

$$\sum_{k=1}^{\infty} p_k = 1 \tag{11}$$

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

$$p_k = \frac{N_k}{N} \tag{12}$$

where N_k is the number of nodes having degree. Hence the number of nodes having degree k can be obtained from the degree distribution as

$$N_k = Np_k. \tag{13}$$

Degree distribution

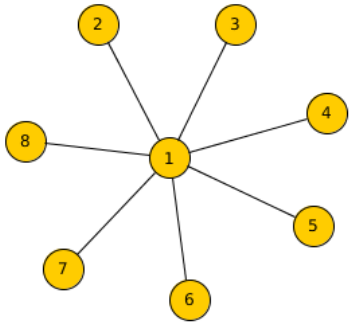
The degree distribution is the most fundamental topological characterisation of a network. It assumed a central role in network theory following the discovery of scale-free networks.

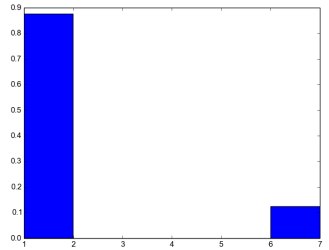
One reason is that the calculation of most network properties requires us to know pk. For example, the average degree of a network can be written as

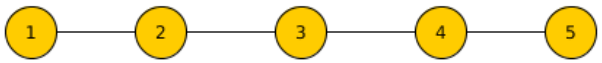
$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k \tag{14}$$

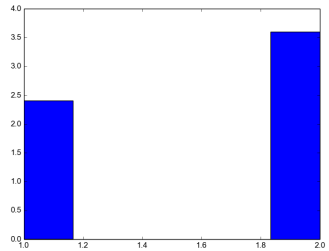
The other reason is that the precise functional form of pk determines many network phenomena, from network robustness to the spread of viruses.

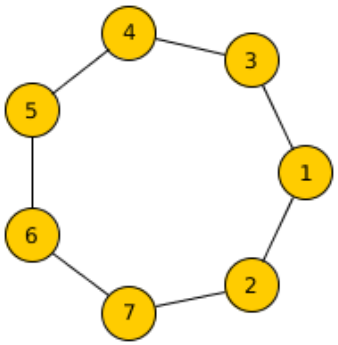
Degree distribution

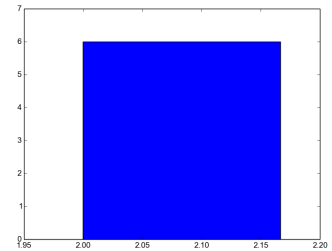












Adjacency matrix

The adjacency matrix of undirected network $G = (V, E)$ (often called simple finite network), the adjacency matrix A is an $n \times n$ matrix with entries such that

$$A_{ij} = \begin{cases} 1 & \text{if there is an edge between vertices } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The diagonal entries of A are zeroes and if the network is undirected, A is a symmetric matrix.

For multigraphs and graphs with loops, the entries are the number of edges between each pair of vertices and the diagonal entries are non-zero due to self-loops which may be counted once or twice based on whether the network is directed or undirected.

Adjacency matrix

For undirected networks a node's degree is a sum over either the rows or the columns of the matrix, i.e.

$$k_i = \sum_{j=1}^{N} A_{ij} = \sum_{i=1}^{N} A_{ji} \tag{16}$$

For directed networks the sums over the adjacency matrix' rows and columns provide the incoming and outgoing degrees, respectively

$$k_i^{in} = \sum_{j=1}^{N} A_{ij}, \quad k_i^{out} = \sum_{j=1}^{N} A_{ji} \tag{17}$$

Given that in an undirected networks the number of outgoing links equals the number of incoming links, we have

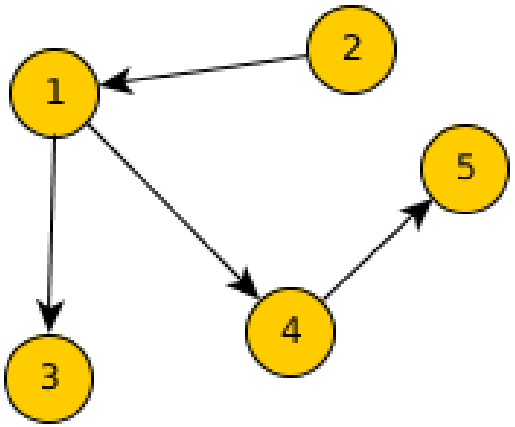
$$2L = \sum_{j=1}^{N} k_i^{in} = \sum_{j=1}^{N} k_i^{out} = \sum_{j=1}^{N} A_{ij} \tag{18}$$

Adjacency matrix

T

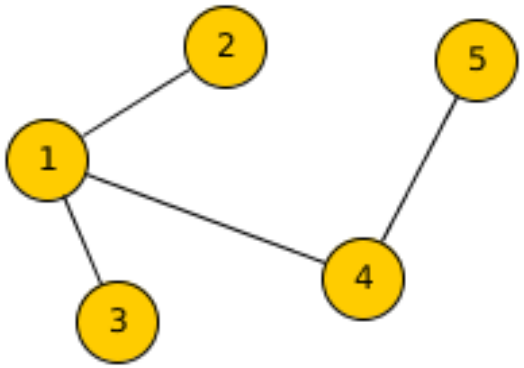
he number of nonzero elements of the adjacency matrix is $2L$, or twice the number of links. Indeed, an undirected link connecting nodes i and j appears in two entries: $A_{ij} = 1$, a link pointing from node j to node i, and $A_{ji} = 1$, a link pointing from i to j.

Adjacency matrix



1. $A_{ij} = ?$
2. $k_2^{in} = ?$
3. $k_2^{out} = ?$
4. $L = ?$
5. $\langle k^{in} \rangle = ?$
6. $\langle k^{out} \rangle = ?$

Adjacency matrix



1. $A_{ij} = ?$
2. $k_2 = ?$
3. $k_2 = ?$
4. $L = ?$
5. $\langle k \rangle = ?$

Sparsity of real networks

In real networks the number of nodes (N) and links (L) can vary widely. For example:

1. The neural network of the worm C. elegans, has $N = 302$ neurons (nodes). (this the only fully mapped nervous system of a living organism).
2. The human brain is estimated to have about a hundred billion ($N \sim 10^{11}$) neurons.
3. The genetic network of a human cell has about $20,000$ genes as nodes;
4. WWW is estimated to have over a trillion web documents ($N > 10^{12}$).

The number of links in a network can also varies widely, between $L = 0$ (null graph, without edges/links) and

$$L_{max} = \frac{N(N-1)}{2} \tag{19}$$

in a complete graph having N nodes.

Sparsity of real networks

- The number of link L in real networks is much smaller than L_{max}, reflecting the fact that most real networks are sparse.
- A network is sparse if $L \ll L_{max}$. For example, This is true for all of the networks in described earlier. One can check that their number of links is only a tiny fraction of the expected number of links for a complete graph of the same number of nodes.

WWW graph

The WWW graph has about 1.5 million links. Yet, if the WWW were to be a complete graph, it should have $L_{max} \approx 5 \times 10^{10}$ links according. Consequently the web graph has only a 3×10^{-5} fraction of the links it could have.

Sparsity of real networks

1. The sparsity of real networks implies that the adjacency matrices are also sparse.
2. Indeed, a complete network has $A_{ij} = 1$, for all (i, j), i.e. each of its matrix elements are equal to one.
3. In contrast in real networks only a tiny fraction of the matrix elements are nonzero.

show the adjacency matrix of C. elangs network and for a complete network having the same number of nodes.

Sparsity of real networks

How we store very large network

Sparseness has important consequences on the way we explore and store real networks. For example, when we store a large network in our computer, it is better to store only the list of links (i.e. elements for which $A_{ij} \neq 0$), rather than the full adjacency matrix, as an overwhelming fraction of the A_{ij} elements are zero. Hence the matrix representation will block a huge chunk of memory, filled mainly with zero.

This can be applied to facebook network for example.

Weighted networks

These are networks in which each link (i, j) has a unique weight w_{ij}. The adjacency matrix of a weighted networks can be defined as follow:

$$A_{ij} = \begin{cases} w_{ij} & \text{if there is an edge between vertices } v_i \text{ and } v_j, \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

1. In mobile call networks the weight can represent the total number of minutes two individuals talk with each other on the phone
2. On the power grid the weight is the amount of current flowing through a transmission line.
3. Most networks of scientific interest are weighted, but we can not always measure the appropriate weights.
4. Consequently we often approximate these networks with an unweighted graph.
5. We will be focusing on unweighted networks, but whenever appropriate, we discuss how the weights alter the corresponding network property

Bipartite networks

1

A network $G = (V, E)$ is bipartite if the nodes can be divided into disjoint sets $V_1 \cup V_2$ such that $(u, v) \in E$ implies that $u \in V_i$, $v \in V_j$, $i \neq j$.

In other words, if we color the V_1-nodes yellow and the V_2-nodes red, then each link must connect nodes of different colors.

2

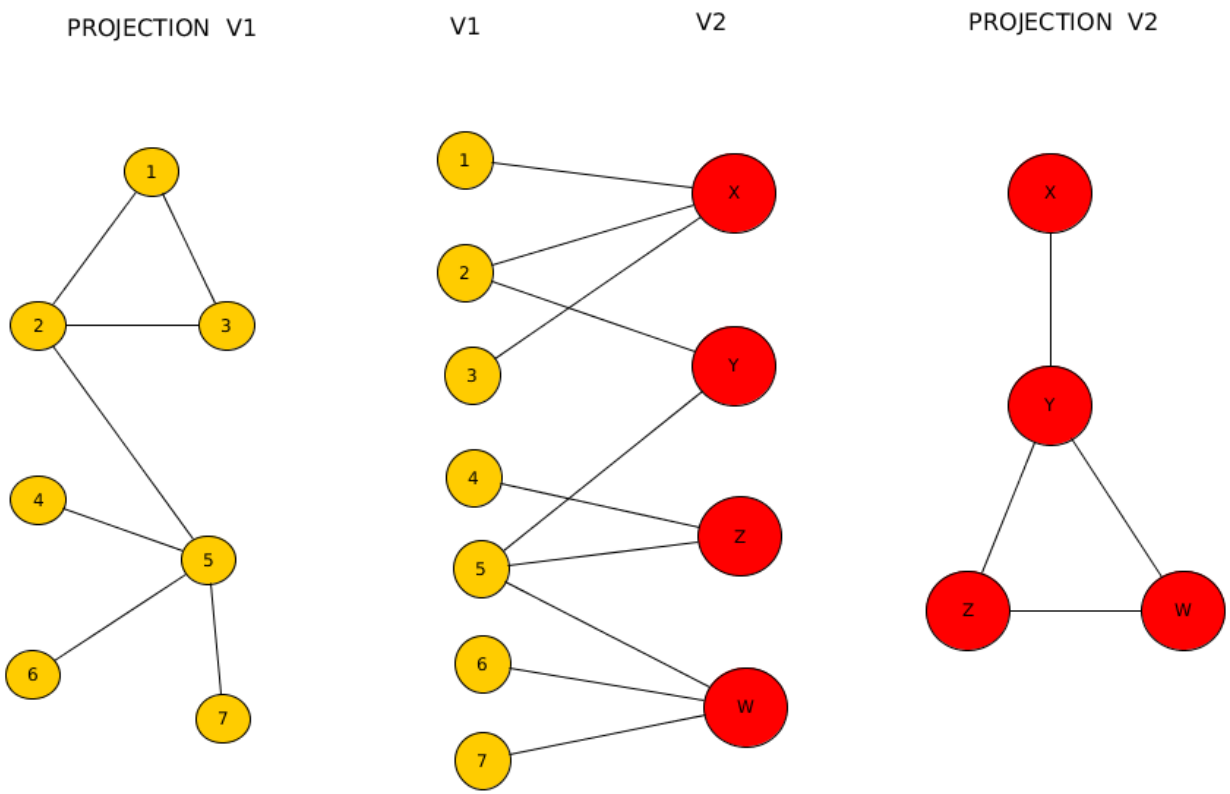
A bipartite network in which each node of V_1 is connected to each node of V_2 is known as a complete bipartite network; if $|V_1| = m$ and $|V_2| = n$, such a network is denoted by $K_{m,n}$.

example

An example of a bipartite network is student-course relation in which we have a set of students and a set of courses. The edges represent courses that a particular student offers.

Bipartite networks

We can generate two projections for each bipartite network. The first projection connects two V_1-nodes by a link if they are linked to the same V_2-node in the bipartite representation. The second projection connects the V_2-nodes by a link if they connect to the same V_1-node.



Bipartite networks

.

In network theory we encounter numerous bipartite networks. A wellknown example is the Hollywood actor network, in which one set of nodes corresponds to movies V_1, and the other to actors V_2.

actor network

A movie is connected to an actor if the actor plays in that movie. One projection of this bipartite network is the actor network, in which two nodes are connected to each other if they played in the same movie.

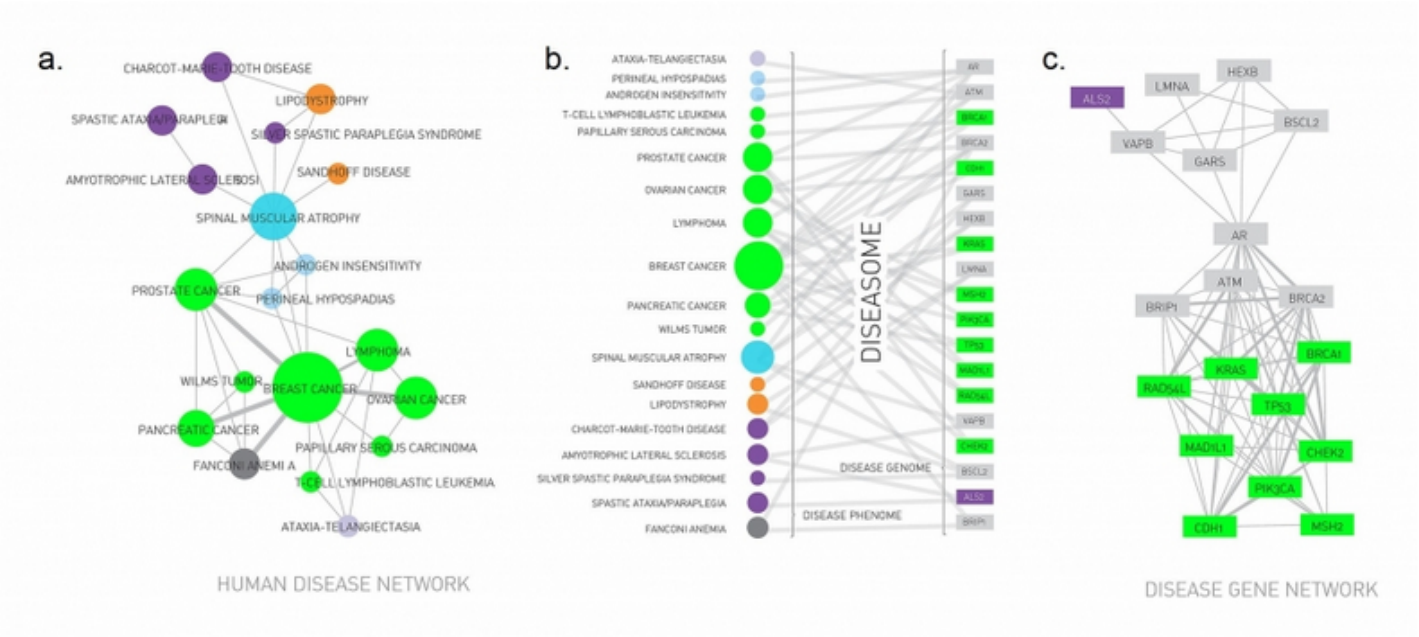
movie network

The other projection is the movie network, in which two movies are connected if they share at least one actor in their cast.

Bipartite networks

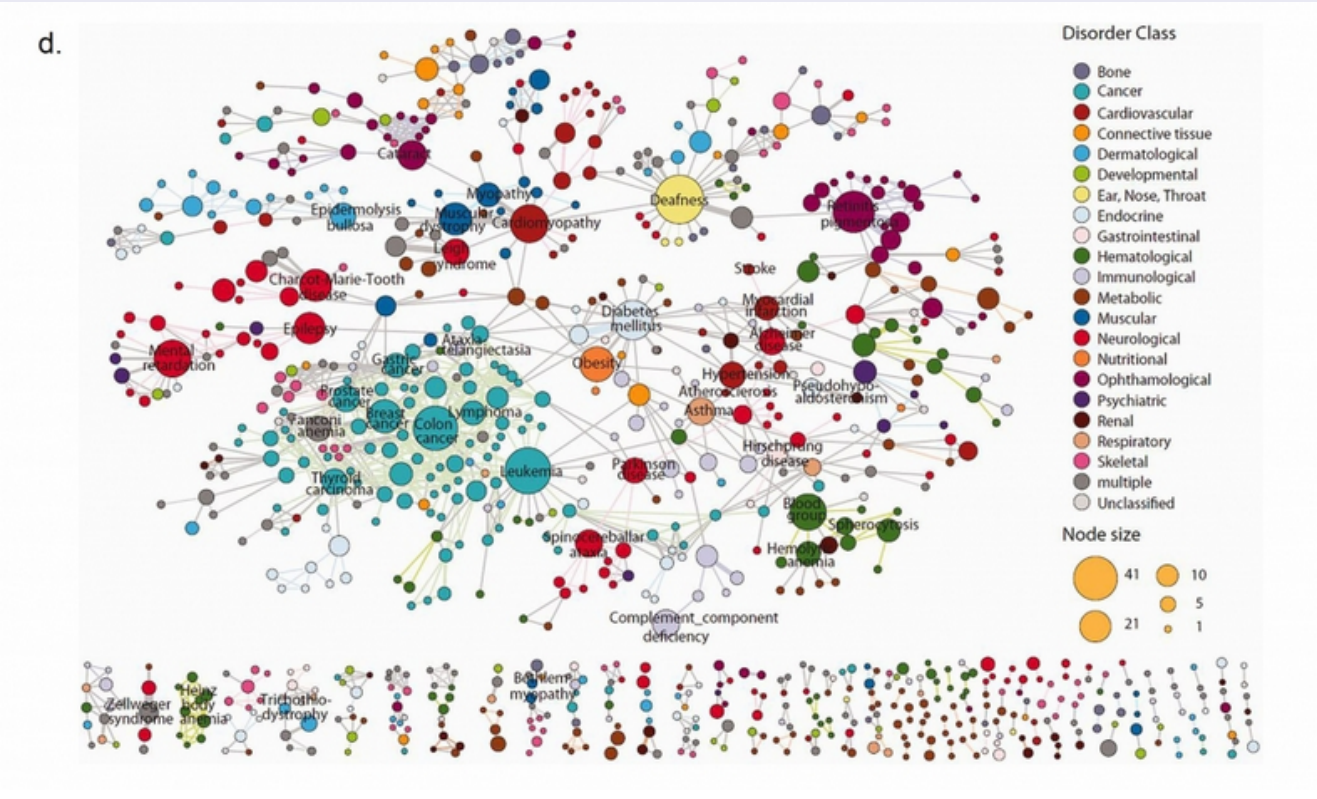
Human Disease Network

Medicine offers another prominent example of a bipartite network: The Human Disease Network connects diseases to the genes whose mutations are known to cause or effect the corresponding disease



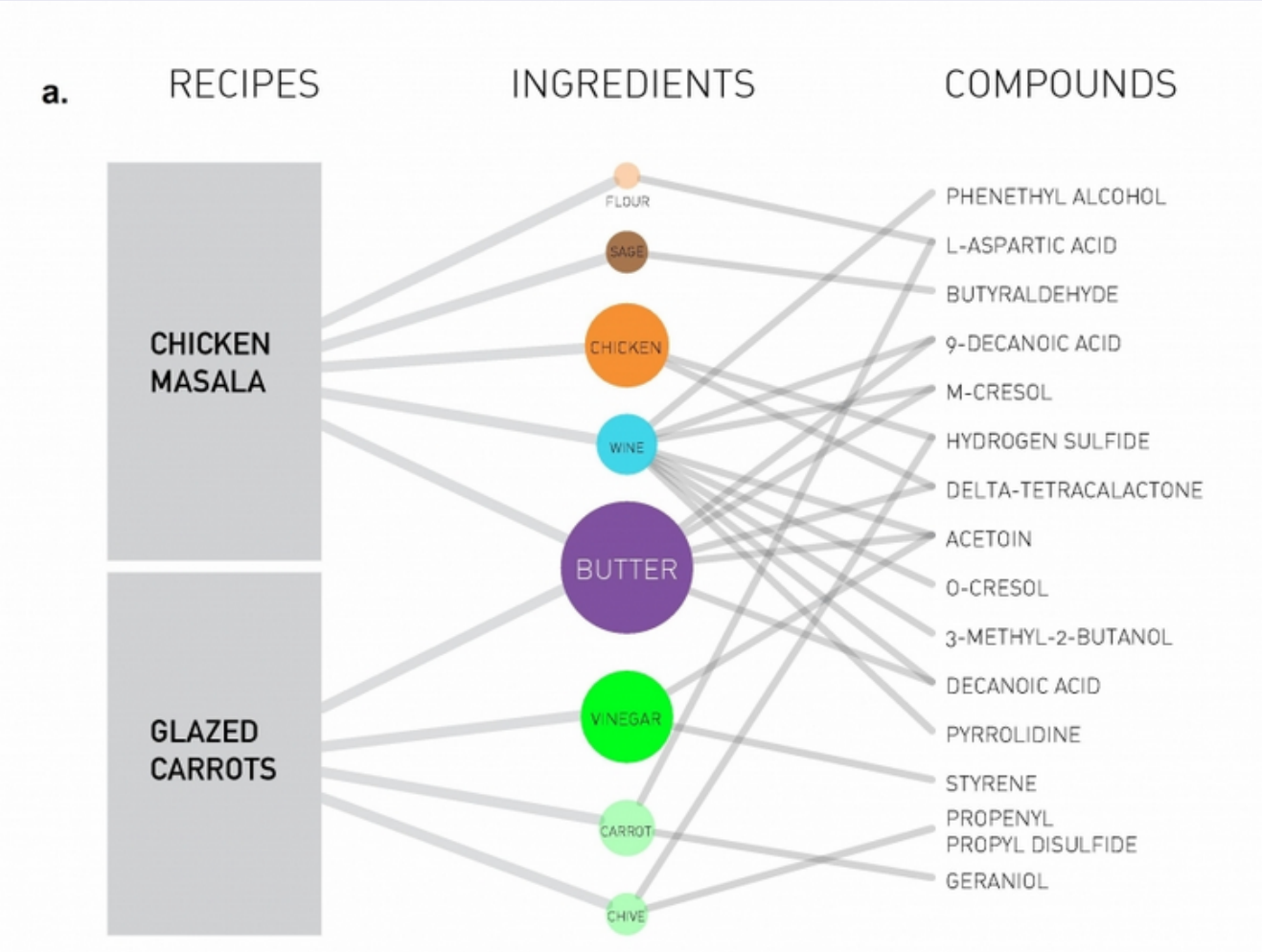
Bipartite networks

Human Disease Network



Multipartite networks

recipe-ingredient-compound



Path and distance

Through-space distance and through links separation

The notion of how far apart two objects are in a physical system intuitively motivates the concept of through-space distance. In a discrete object, however, this concept involves the through-links separation of two nodes in the network.

network metric

Is a function which defines a distance between the nodes of the network, such that $m : V(G) \times V(G) \to \mathbb{R}^+$ such that:

1. $m(u, v) \geq 0 \ \forall u \in V, v \in V$.
2. $m(u, v) = 0$ if and only if $u = v \ \forall u \in V, v \in V$.
3. $m(u, v) = m(v, u) \ \forall u \in V, v \in V$.
4. $m(u, w) = m(u, v) + m(v, w) \ \forall u \in V, v \in V, w \in V$.

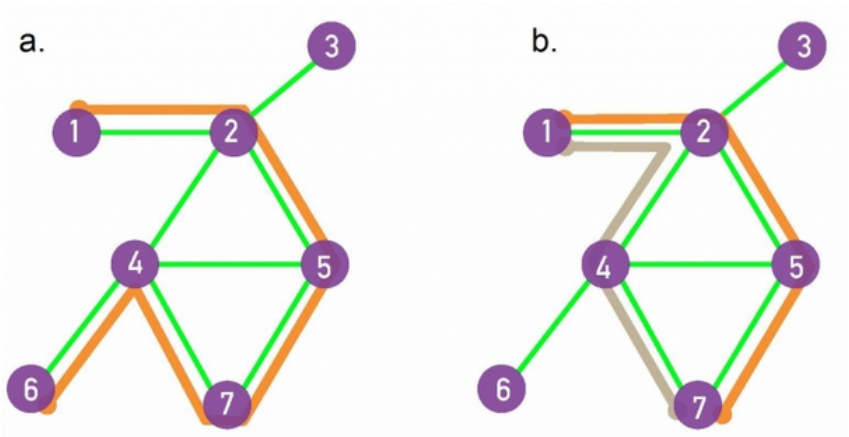
Path, Shortest path

.

In networks physical distance is replaced by path length. A path is a route that runs along the links of the network. A path's length represents the number of links the path contains.

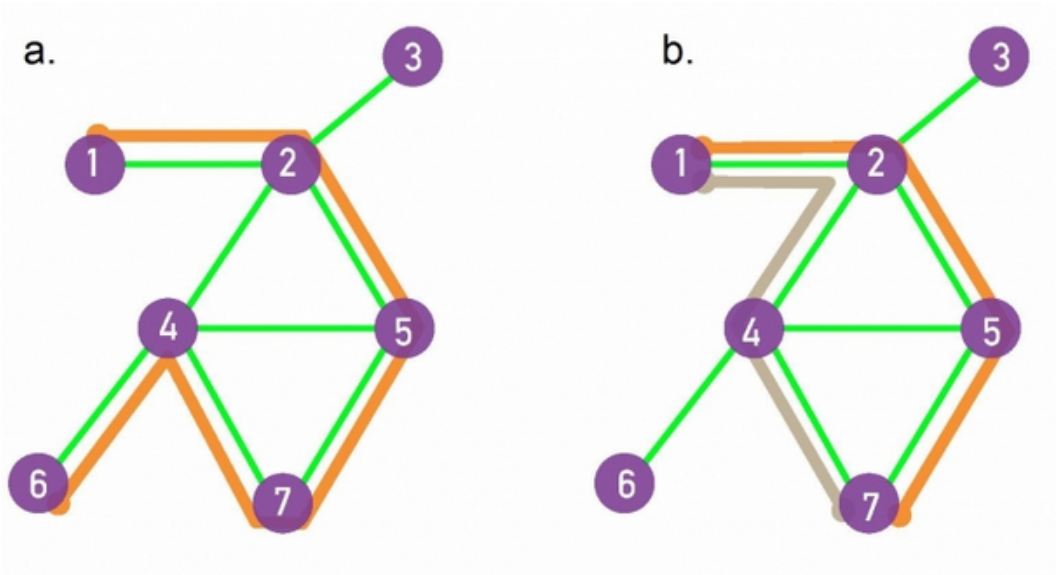
path

A path between nodes i_0 and in is an ordered list of n links $P = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \cdots, (i_{n-1}, i_n)\}$. The length of this path is n. The path shown in orange in (a) follows the route $1 \to 2 \to 5 \to 7 \to 4 \to 6$, hence its length is $n = 5$.



Shortest Path

Although there are some alternative ways of defining the distance between two nodes in a network, the notion of shortest path distance is widely recognised as the standard definition of network distance.

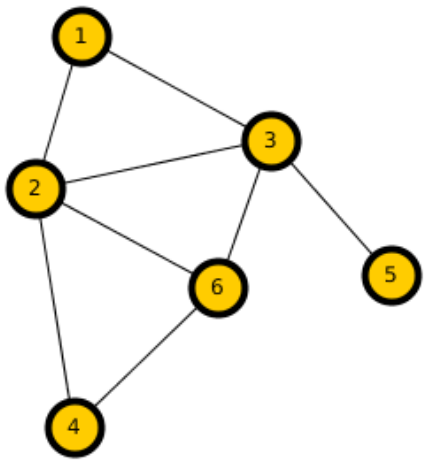


.

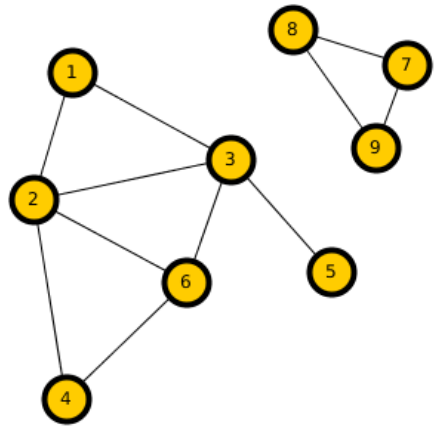
The shortest paths between nodes 1 and 7, or the distance d_{17}, correspond to the path with the fewest number of links that connect nodes 1 to 7. There can be multiple paths of the same length, as illustrated by the two paths shown in orange and grey.

Shortest Path: undirected network

In an undirected network, the shortest path distance $d(u, v)$ is the number of links in the shortest path between the nodes u and v in a network. In the case when u and v are in different connected components of the network, the distance between them is set to infinite, $d(u, v) = \infty$.



$d(2, 3) = 1,\ d(5, 6) = 2,$
$d(1, 6) = 1$



$d(5, 6) = 2,\ d(1, 6) = 2,$
$d(3, 8) = \infty,\ d(5, 7) = \infty$

The number of shortest paths, N_{ij}, and the distance d_{ij} between nodes i and j can be calculated directly from the adjacency matrix A_{ij}.

1. $d_{ij} = 1$: If there is a direct link between i and j, then $A_{ij} = 1$ ($A_{ij} = 0$ otherwise).
2. $d_{ij} = 2$: If there is a path of length two between i and j, then $A_{ik} \cdot A_{kj} = 1$ ($A_{ik} \cdot A_{kj} = 0$ otherwise).
3. The number of $d_{ij} = 2$ paths between i and j is

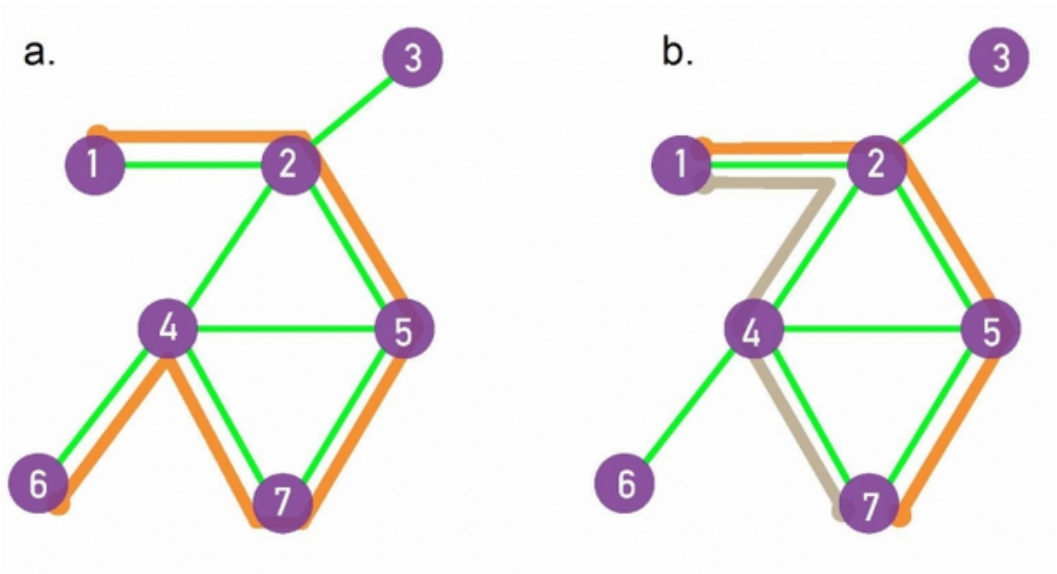
$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik} \cdot A_{kj} = A_{ij}^2 \tag{21}$$

4. The number of paths of length d between i and j is

$$N^{(d)} = A_{ij}^d \tag{22}$$

$d_{ij} = d$: If there is a path of length d between i and j, then $A_{ik} \cdots A_{kj} = 1$ ($A_{ik} \cdots A_{kj} = 0$ otherwise).

In real networks we often need to determine the distance between two nodes. For a small network, this is an easy task. For a network with millions of nodes finding the shortest path between two nodes can be rather time consuming. The length of the shortest path and the number of such paths can be formally obtained from the adjacency matrix. In practice the breadth first search (BFS) algorithm is used to find the distance between two nodes.

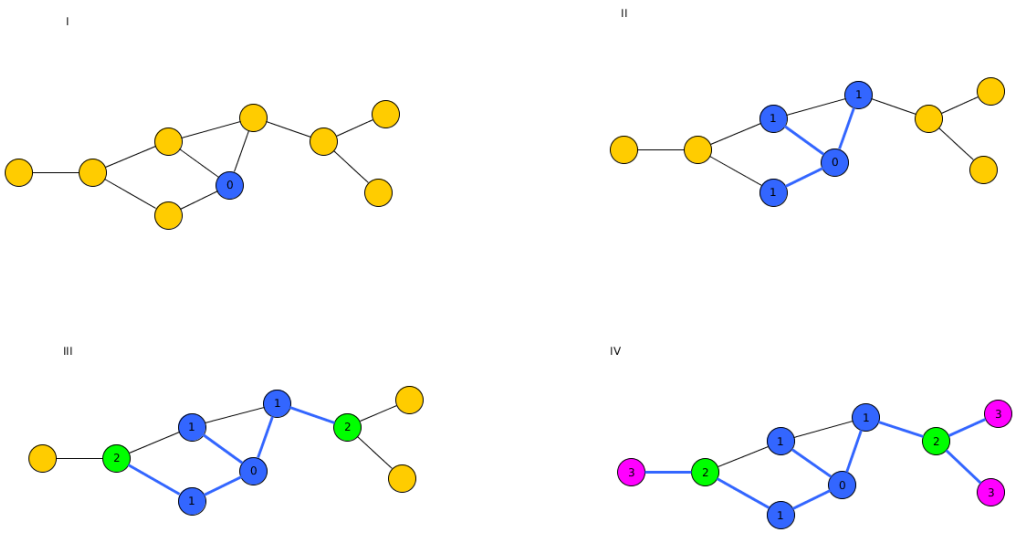


Shortest Path:Breadth-First Search (BFS) Algorithm

BFS starts from a node and labels its neighbors, then the neighbors' neighbors, until it reaches the target node. The number of "ripples" needed to reach the target provides the distance.
The identification of the shortest path between node i and j follows the following steps

1. Start at node i, that we label with "0".
2. Find the nodes directly linked to i. Label them distance "1" and put them in a queue.
3. Take the first node, labelled n, out of the queue ($n = 1$ in the first step).
4. Find the unlabelled nodes adjacent to it in the graph. Label them with $n + 1$ and put them in the queue.
5. Repeat step 3 until you find the target node j or there are no more nodes in the queue.
6. The distance between i and j is the label of j. If j does not have a label, then $d_{ij} = \infty$.

Applying the BFS Algorithm



1. Start from the orange node, labelled 0, we identify all its neighbours, labelling them 1.
2. Next label 2 the unlabelled neighbours of all nodes labelled 1, and so on, in each iteration increasing the label number, until no node is left unlabelled.
3. The length of the shortest path or the distance d_{0i} between node 0 and any other node i in the network is given by the label of node i.
4. For example, the distance between node 0 and the leftmost node is $d = 3$.

Pathology

Cycle

A path with the same start and end node. In the graph shown on the left we have only one cycle, as shown by the orange line.

Eulerian path

A path that traverses each link exactly once. The image shows two such Eulerian paths, one in orange and the other in blue.

Eulerian path

A path that visits each node exactly once. We show two Hamiltonian paths in orange and in blue.

Shortest Path: directed network

In directed network, the directed distance $\overrightarrow{d}(u, v)$ between a pair of nodes u and v in a directed network is considered to be the length of the directed shortest path from u to v. When there is no a directed path connecting two nodes, the corresponding distance is infinite.

pseudo-distance

in general for directed network we have $\overrightarrow{d}(u, v) = \overrightarrow{d}(v, u)$, which violates the symmetry property of metric functions. Therefore, $\overrightarrow{d}(u, v)$ is not a distance but a pseudo-distance or pseudo-metric.

Strongly connected directed network

A directed network is referred to as **strongly connected** if there is a directed path from u to v and a directed path from v to u for every pair of distinct nodes in the network.

Distance matrix

distance matrix

In an undirected connected network the distance matrix **D** is such the entries are the distance between any pair of node in the network. The distance matrix is square and the distances between a node v and any other node in the network are given at the vth row or column of **D**.
In the case of a directed network, the distance matrix is not necessarily symmetric and can contain entries equal to infinite.

eccentricity

The maximum entry for a given row/column of the distance matrix of an undirected (strongly connected directed) network is known as the eccentricity $e(u)$ of the node u and given by:

$$e(u) = \max_{v \in V(G)} \{d(u, v)\} \tag{23}$$

Distance matrix

diameter

The maximum eccentricity among the nodes of a network is known as the diameter of the network, which is given by:

$$diam(G) = \max_{u,v \in V(G)} \{d(u,v)\} \tag{24}$$

radius

The radius of the network is the minimum eccentricity of the nodes, and a node is called central if its eccentricity is equal to the radius of the network. the centre of the graph $\mathcal{C}(G)$ is the set of all central nodes.

$$r(G) = \min_{u,v \in V(G)} \{d(u,v)\}, \tag{25}$$

$$\mathcal{C}(G) = \{u \in V(G) | \text{u is central node}\}. \tag{26}$$

Distance

sum distance

The sum of all entries of a row/column of the distance matrix is known as the distance sum $s(u)$ of the corresponding node

$$s(u) = \sum_{u \in V(G)} d(u, v) \tag{27}$$

This sum is also called the **total distance** or **status** of the node.

Weiner index

This index is defined as the semi-sum of all entries of the distance matrix and was introduced by Weiner in 1947 to to account for the variations of molecular branching in hydro carbons and is given by

$$W(G) = \frac{1}{2} \sum_{u} \sum_{v} d(u, v) = \sum_{u=1}^{N} s(u) = \frac{1}{2} \mathbf{1}^T \mathbf{D} \mathbf{1}. \tag{28}$$

Average path length

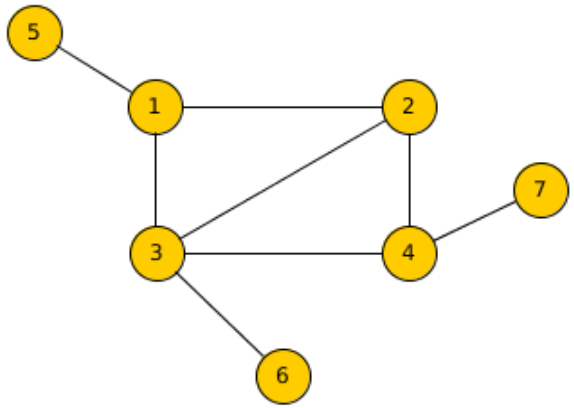
An important characteristic descriptor of the topology of a network is its average path length which can be expressed in terms of the Wiener index and given by:

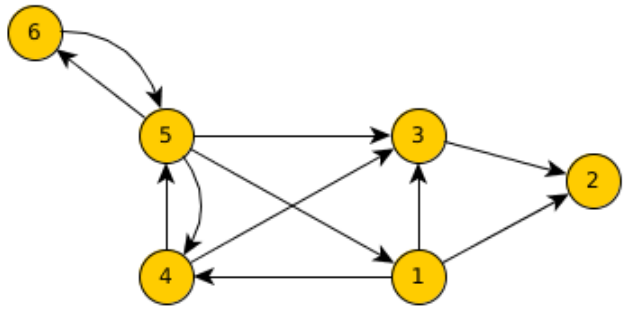
$$\bar{l} = \frac{2W(G)}{n(n-1)} \tag{29}$$

where the average is taken by considering only those pairs of nodes for which a path connecting them exists.

Example

Compute the distance, eccentricity, radius, centre, Weiner index for each of the following network





connectedness

.

One of the most they utility of most networks is to ensure connectedness. For example a phone would be of limited use as a communication device if we could not call any valid phone number; email would be rather useless if we could send emails to only certain email addresses, and not to others.

.

From a network science perspective, the network behind the phone or the Internet must be capable of establishing a path between any two nodes.

.

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 $d_{ij} = \infty$.

connected network

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 $d_{ij} = \infty$. Clearly the network shown in Image 2.15a is disconnected, and we call its two subnetworks components or clusters.

component

A component is a subset of nodes in a network, so that there is a path between any two nodes that belong to the component, but one cannot add any more nodes to it that would have the same property.

bridge

If a network consists of two components, a properly placed single link can connect them, making the network connected. Such a link is called a bridge. In general a bridge is any link that, if cut, disconnects the network.

Identifying components of large networks

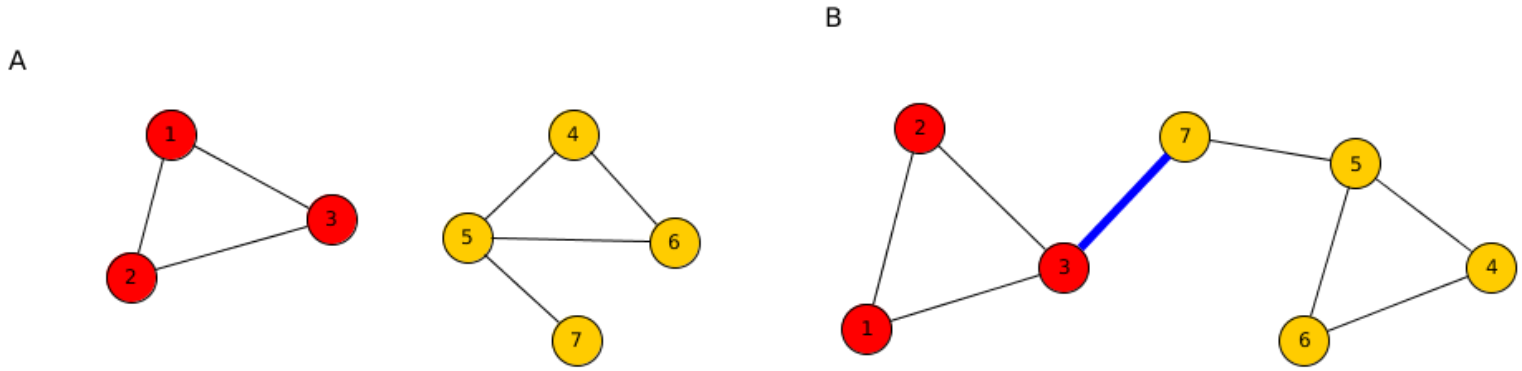
While for a small network visual inspection can help us decide if it is connected or disconnected, for a network consisting of millions of nodes connectedness is a challenging question.

Mathematical and algorithmic tools can help us identify the connected components of a graph.

example

for a disconnected network the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements in the matrix are contained in square blocks along the matrix' diagonal and all other elements are zero.

Each square block corresponds to a component. We can use the tools of linear algebra to decide if the adjacency matrix is block diagonal, helping us to identify the connected components.



(a) disconnected network (b) connected network

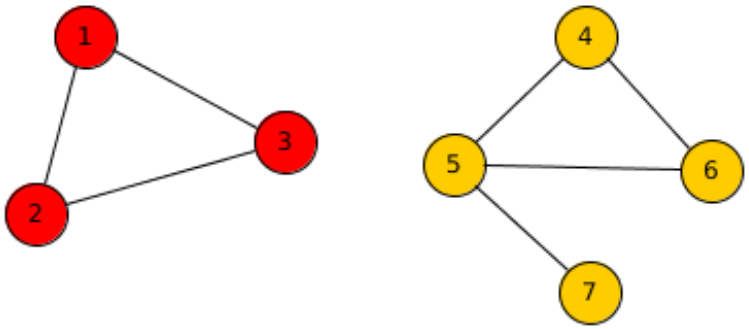
Disconnected components

A small network consisting of two disconnected components. Indeed, there is a path between any pair of nodes in the $(1, 2, 3)$ component, as well in the $(4, 5, 6, 7)$ component. However, there are no paths between nodes that belong to the different components.

The addition of a single link, called a bridge, shown in red, turns a disconnected network into a single connected component. Now there is a path between every pair of nodes in the network. Consequently the adjacency matrix cannot be written in a block diagonal form.

connected and disconnected networks

A

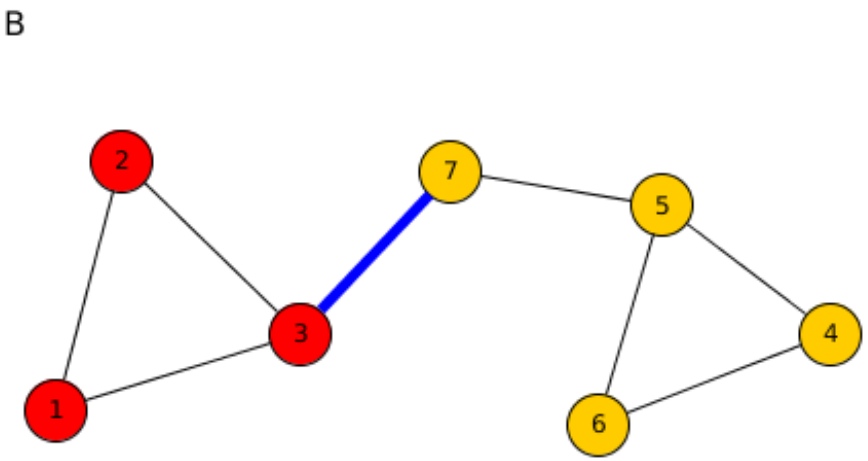


$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix},$$

connected components

If the network has disconnected components, the adjacency matrix can be rearranged into a block diagonal form, such that all nonzero elements of the matrix are contained in square blocks along the diagonal of the matrix and all other elements are zero.

connected and disconnected networks



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

bridge

The addition of a single link, called a bridge, shown in blue, turns a disconnected network into a single connected component. Now there is a path between every pair of nodes in the network. Consequently the adjacency matrix cannot be written in a block diagonal form.

Finding the Connected Components of a Network

1. Start from a randomly chosen node i and perform a BFS. Label all nodes reached this way with $n = 1$.
2. If the total number of labelled nodes equals N, then the network is connected. If the number of labelled nodes is smaller than N, the network consists of several components. To identify them, proceed to step 3.
3. Increase the label $n \to n + 1$. Choose an unmarked node j, label it with n. Use BFS to find all nodes reachable from j, label them all with n. Return to step 2.

Clustering Coefficient: Global and Local

- Clustering, also known as transitivity, is a typical property of acquaintance networks,
- where two individuals with a common friend are likely to know each other,
- In terms of network topology, transitivity means the presence of a high number of triangles,
- Two versions of this measure exist:
  - the global clustering coefficient and
  - the local clustering coefficient

The global version was designed to give an overall indication of the clustering in the network, whereas the local gives an indication of the embeddedness of single nodes.

Global Clustering Coefficient

The global clustering coefficient of a network can be seen as the relative number of transitive triples (expression borrowed from the sociology literature), i.e. the fraction of connected triples of nodes (triads) which also form triangles:

$$C = \frac{3 \times \text{ number of triangles in the network}}{\text{number of connected triplets in the network}}$$

or

$$C = \frac{\text{number of closed triplets}}{\text{number of connected triplets of vertices}}.$$

Global Clustering Coefficient

- A triplet is three nodes that are connected by either two (open triplet) or three (closed triplet) undirected ties.
- A triangle consists of three closed triplets, one centred on each of the nodes.
- The global clustering coefficient is the number of closed triplets (or $3\times$ number of triangles) over the total number of triplets (both open and closed).
- The factor 3 in the numerator compensates for the fact that each complete triangle of three nodes contributes three connected triplets, one centred on each of the three nodes, and ensures that $0 \leq C \leq 1$, with $C = 1$ for the complete graph K_N.

Local Clustering Coefficient: Watts and Strogatz

- Probability that nearest neighbours of a node are themselves nearest neighbours, i.e.
- Concretely if node i has k_i nearest neighbours with e_i connections i.e.
- The value obtained by counting the actual number of edges in G_i (the subgraph of neighbours of i)
- the local clustering coefficient is defined as the ratio between e_i and $k_i(k_i-1)/2$

$$c_i = \frac{e_i}{k_i(k_i-1)/2}.$$

$k_i(k_i-1)/2$ is the maximum possible number of edges in G_i:

For this alternative, the clustering coefficient of a network is defined to be the average of c_i over all the nodes in the network:

$$\overline{C} = \frac{1}{n}\sum_{i=1}^{n} c_i.$$

Local Clustering Coefficient: Watts and Strogatz

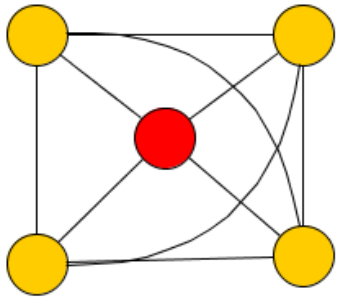
1. $c_i = 0$ if none of the neighbors of node i link to each other.
2. $c_i = 1$ if the neighbors of node i form a complete graph, i.e. they all link to each other.
3. c_i is the probability that two neighbors of a node link to each other. Consequently $c = 0.5$ implies that there is a 50% chance that two neighbors of a node are linked.

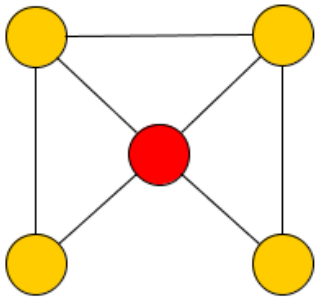
summary

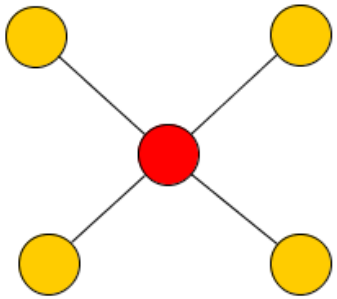
c_i measures the network's local link density: The more densely interconnected the neighborhood of node i, the higher is its local clustering coefficient.

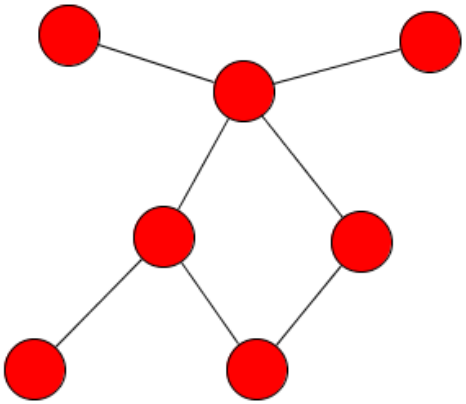
Local Clustering Coefficient

Find the local clustering coefficient of the red node









The End