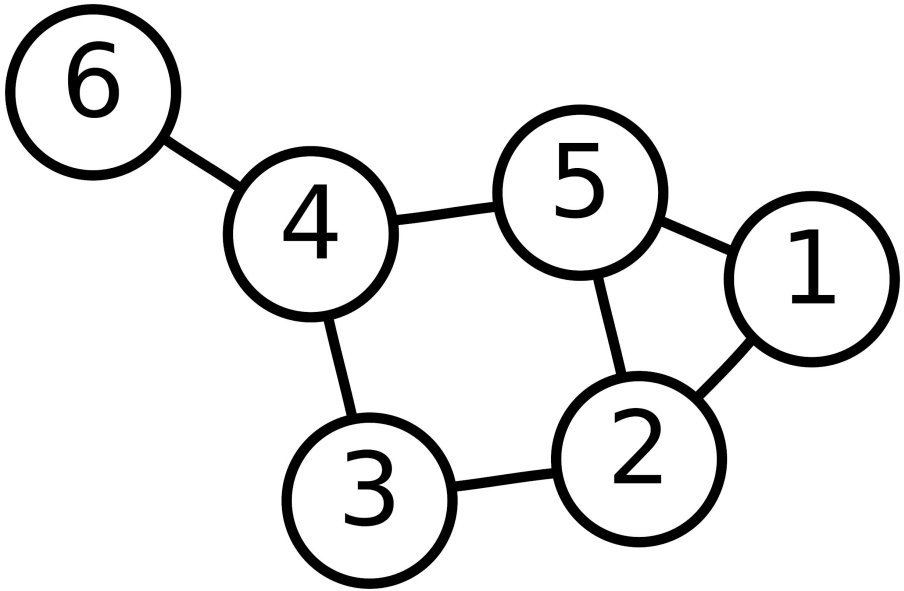
Network analysis

Why networks?

A way of representing data that emphasises **relationships** or **interactions**

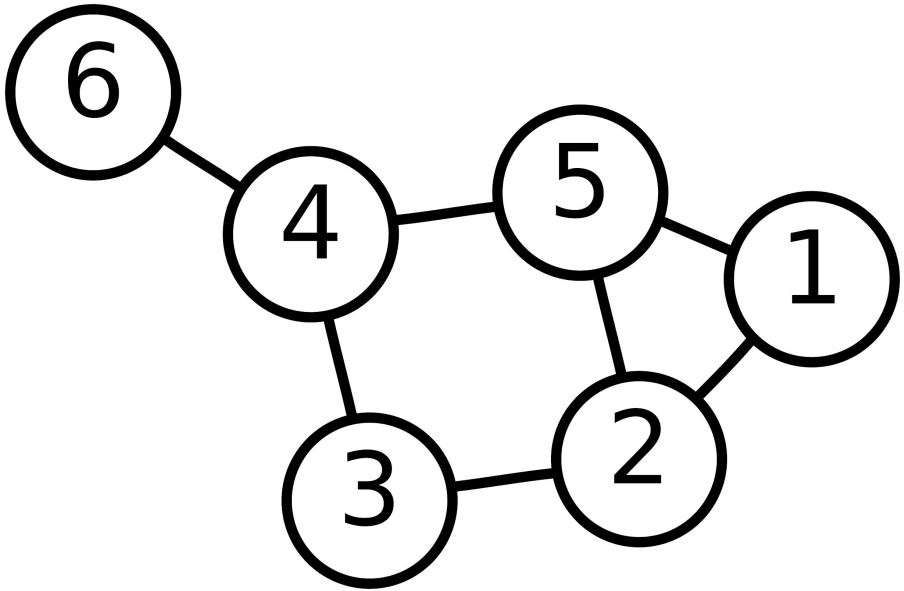


- Protein–protein interaction network
- Gene regulatory networks
- Gene co-expression networks

Network analysis

Network analysis is a natural way of studying those relationships in a structured way.

It provides ways to get both quantitative and qualitative descriptions of the relationships in the data.

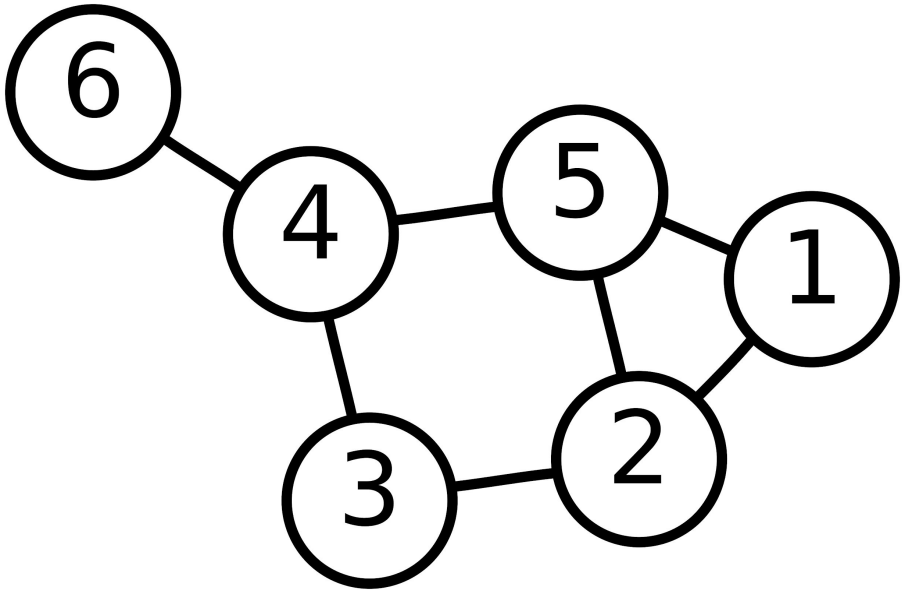


A network

Components: nodes or vertices

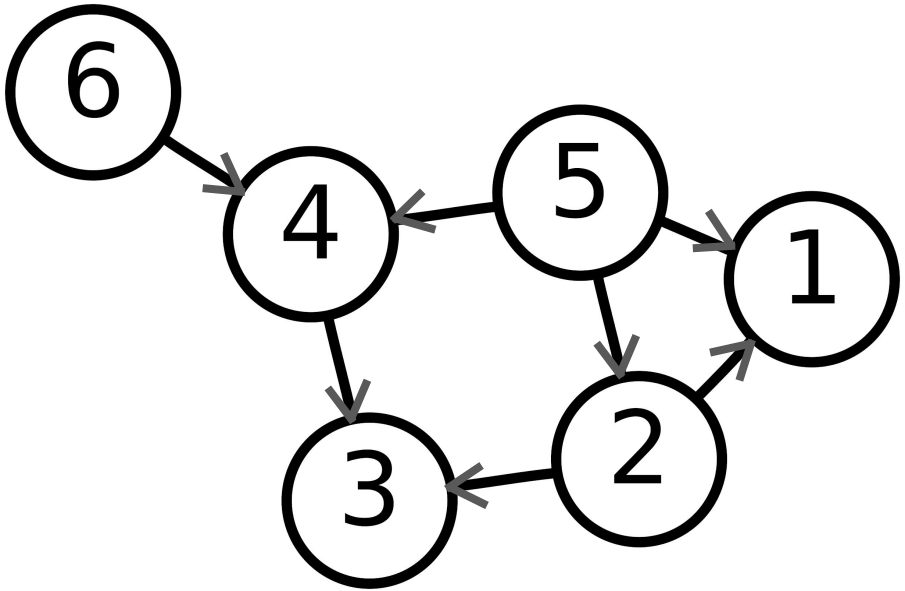
Interactions: links or edges

Together: network or graph



A directed network

If the relationships or the interactions are not symmetric, you may represent it using a directed network.



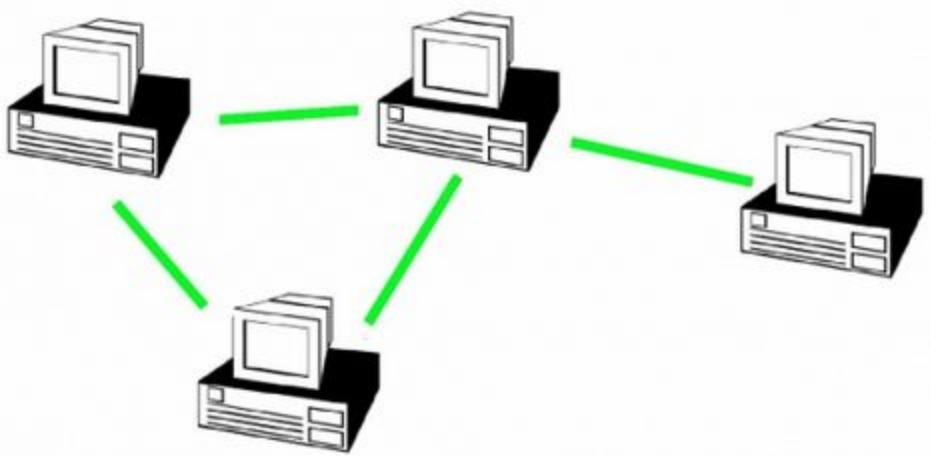
Edge weights

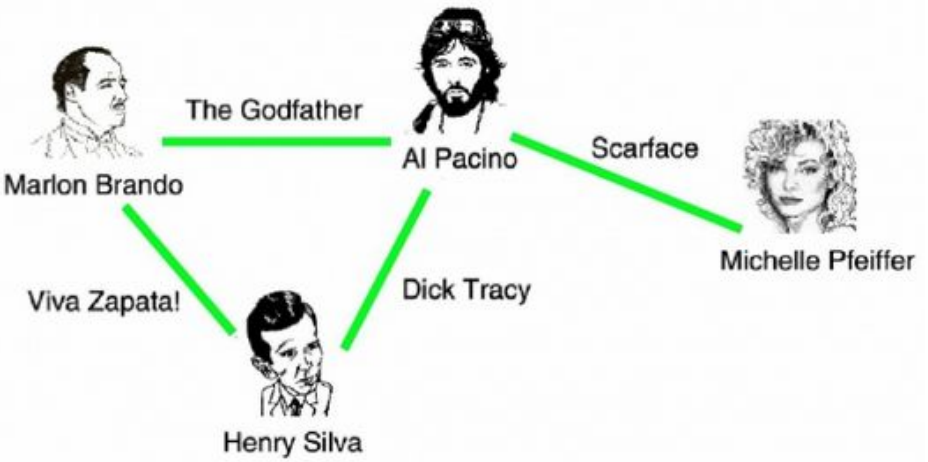
Both nodes and links can contain any number of **attributes**.

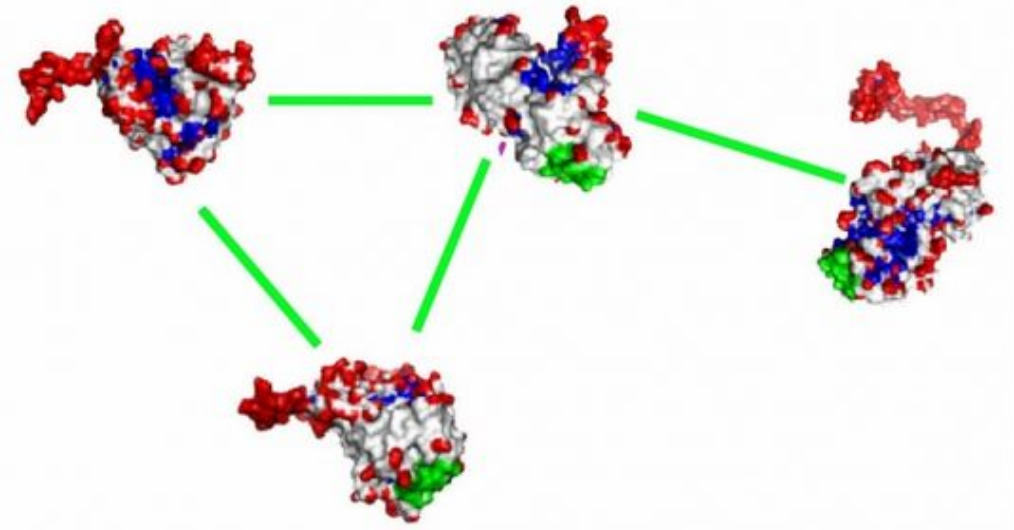
One of the most common and useful edge attributes is the **edge weight**.

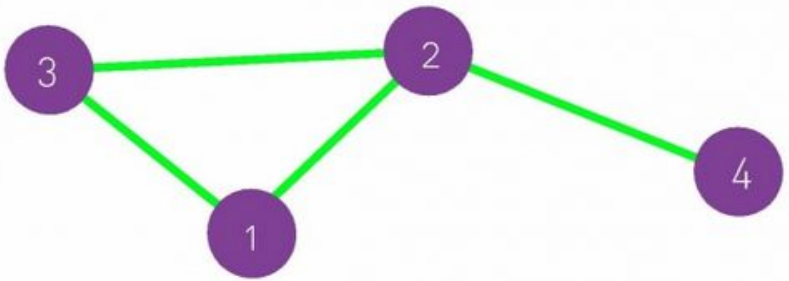
It represents the strength of the interaction between the two nodes.

Network properties




The Godfather
Marlon Brando
Al Pacino
Scarface
Michelle Pfeiffer
Viva Zapata!
Dick Tracy
Henry Silva




3
2
4
1

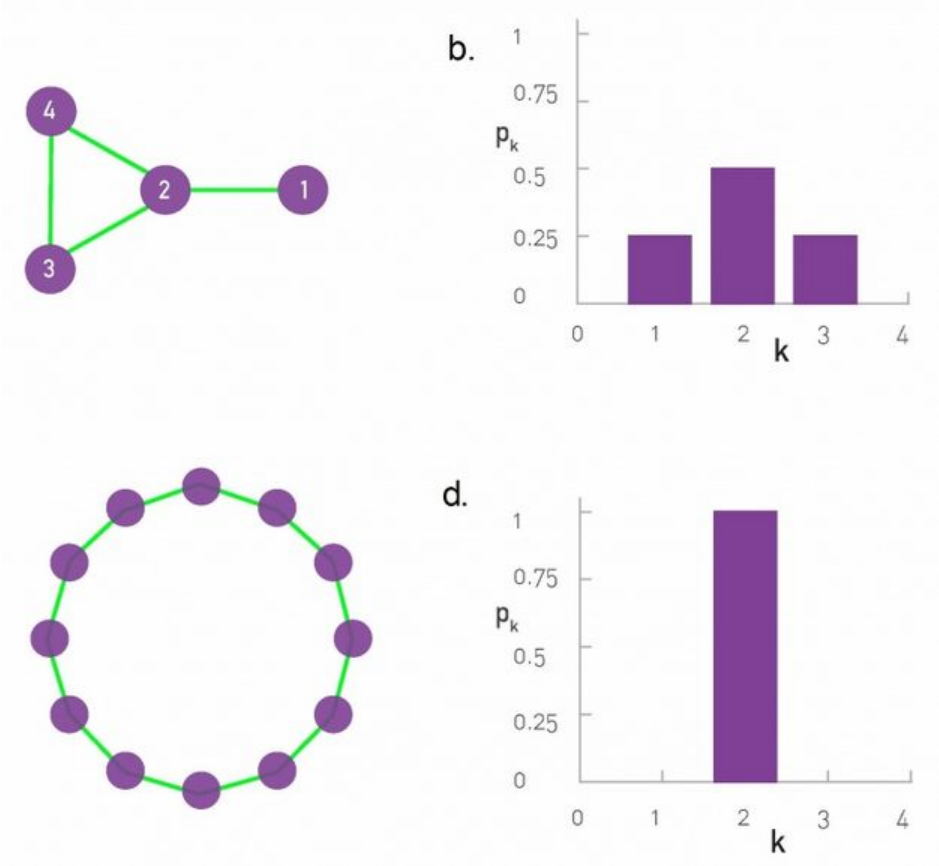
Node degree

The degree of a node is the **number of edges** it has.

In directed networks we have **in degree** and **out degree**.

Degree distribution

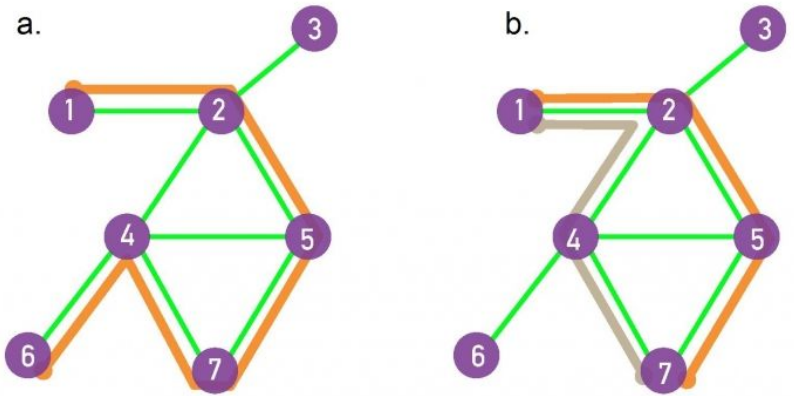
The node degree distribution gives a good glimpse on the structure of the network.

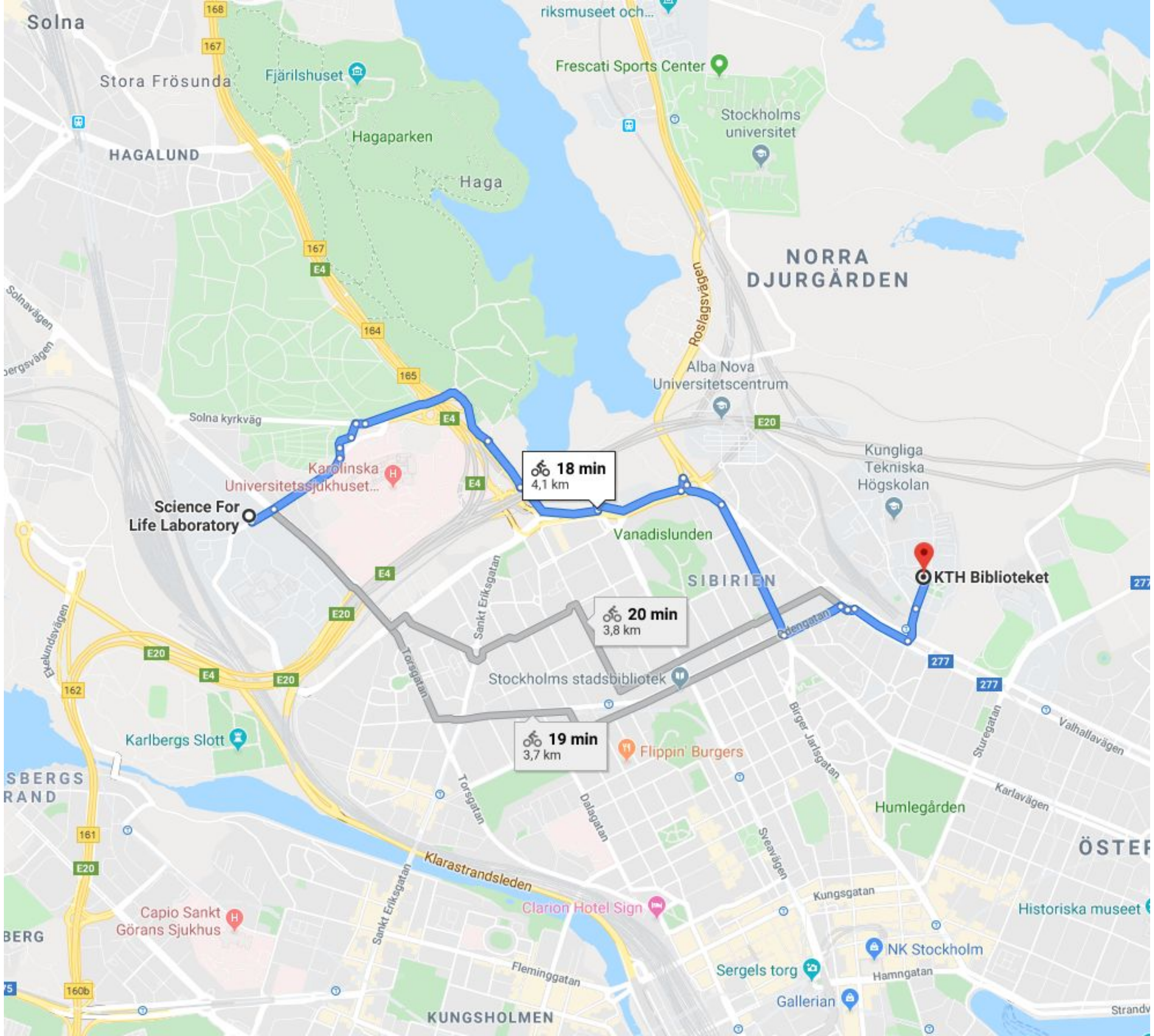


Paths

A **path** is any collection of edges that connect two vertices together. The **length** of the path is the number of edges it contains.

The **shortest path** is the path with the smallest length that connect two edges.




Solna
Stora Frösunda
HAGALUND
Fjärilshuset
Hagaparken
Haga
riksmuseet och...
Frescati Sports Center
Stockholms universitet
NORRA DJURGÅRDEN
Roslagsvägen
Alba Nova Universitetscentrum
Kungliga Tekniska Högskolan
Solna kyrkväg
Karolinska Universitetssjukhuset...
Science For Life Laboratory
18 min
4,1 km
Vanadislunden
SIBIRIEN
KTH Biblioteket
20 min
3,8 km
19 min
3,7 km
Sankt Eriksgatan
Torsgatan
Odengatan
Stockholms stadsbibliotek
Flippin' Burgers
Birger Jarlsgatan
Karlbergs Slott
Dalagatan
Sveavägen
Sturegatan
Valhallavägen
Karlavägen
Humlegården
Klarastrandsleden
Clarion Hotel Sign
Kungsgatan
Capio Sankt Görans Sjukhus
Historiska museet
NK Stockholm
Sergels torg
Hamngatan
Fleminggatan
Gallerian
KUNGSHOLMEN
Ekelundsvägen
Solnavägen

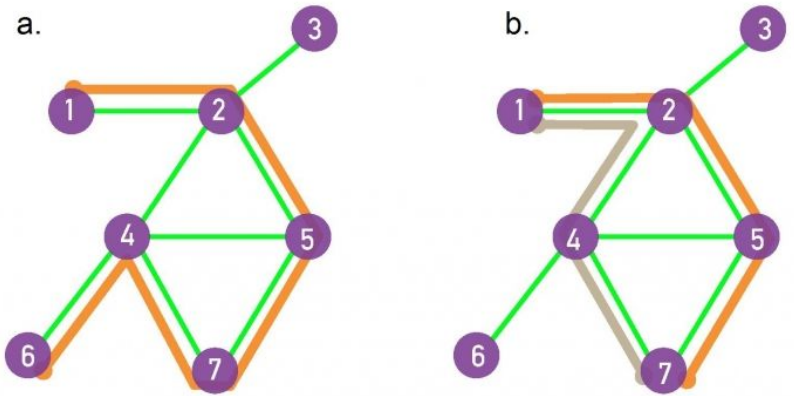
Paths

The **diameter** of the network is the length of its largest shortest path.

The **average shortest path** is a good metric of how connected your network is.

A **cycle** is a path that starts and ends in the same node, without using the same edge or node twice.

A **tree** is a graph with no cycles.



Connected components

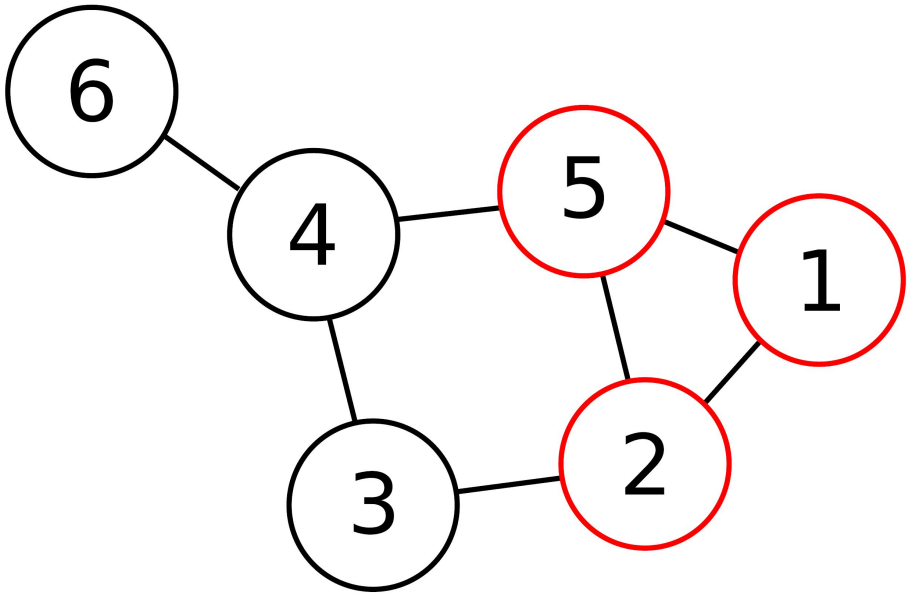
It is not necessary for a path to exist between any two nodes in a network.

A **connected component** of a network is a subset of nodes (and its links) where there exists a path between any two nodes.

Maximum clique

A **clique** is a set of nodes where all the nodes are connected to every other nodes

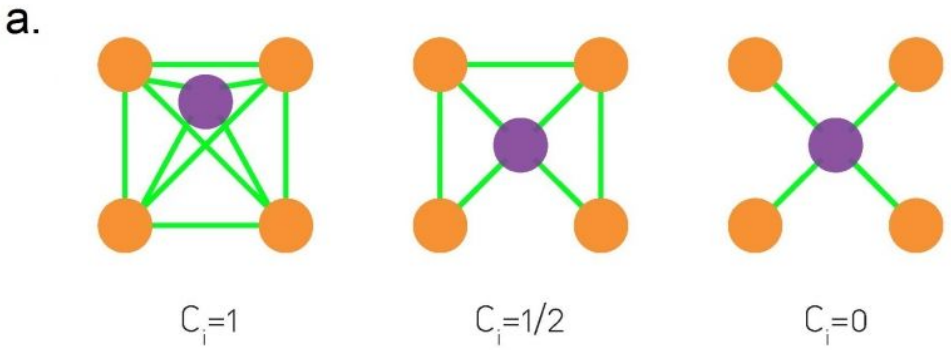
Since every subset of a clique is also a clique, we talk of **maximum cliques** as cliques that are not a subset of any larger clique



Clustering coefficient

The **clustering coefficient** is a node specific measure of how it's neighbours are connected to each other.

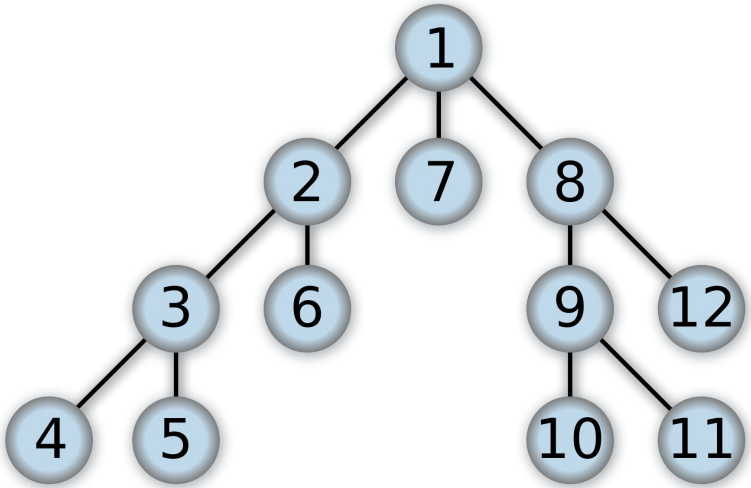
The coefficient is **1** if all the neighbours of a node connect to each other, and **0** if there are no connections between neighbours.



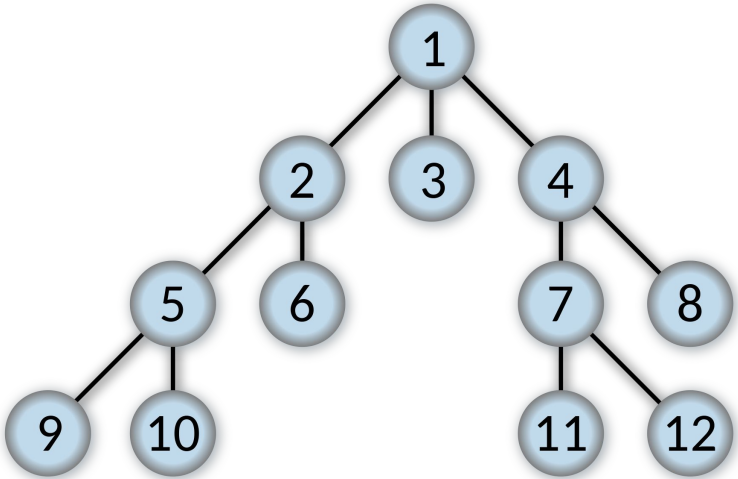
Graph traversal algorithms

A structured way of visiting all the nodes in a graph.

Depth-first search

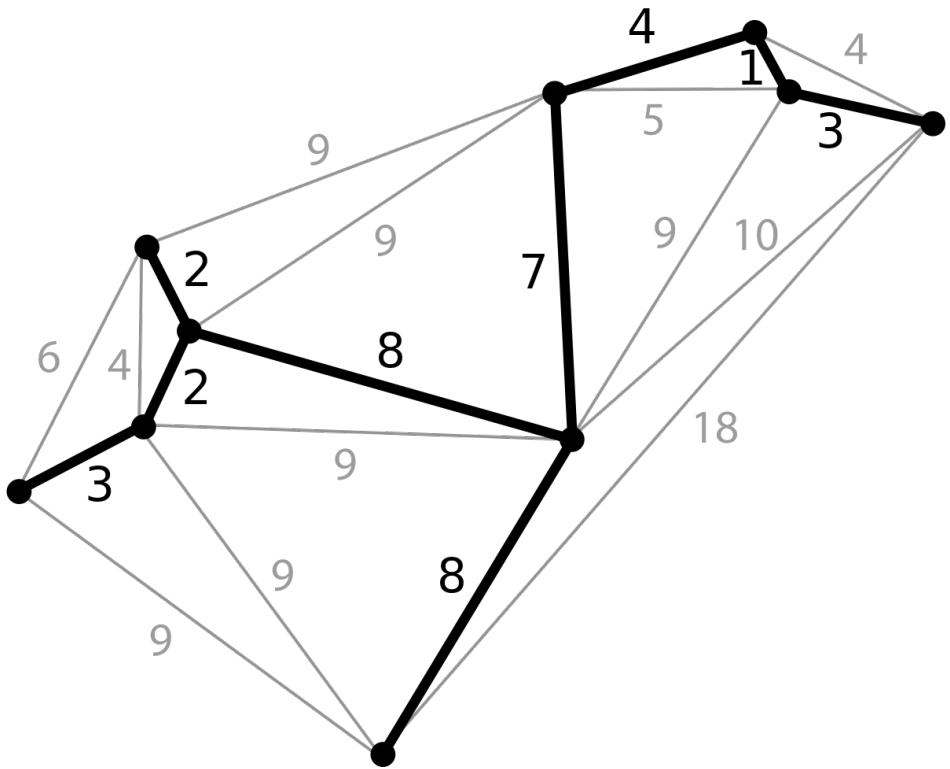


Breadth-first search



Minimum spanning tree

The **minimum spanning tree** connects all the nodes in a graph with the minimum amount of links possible.

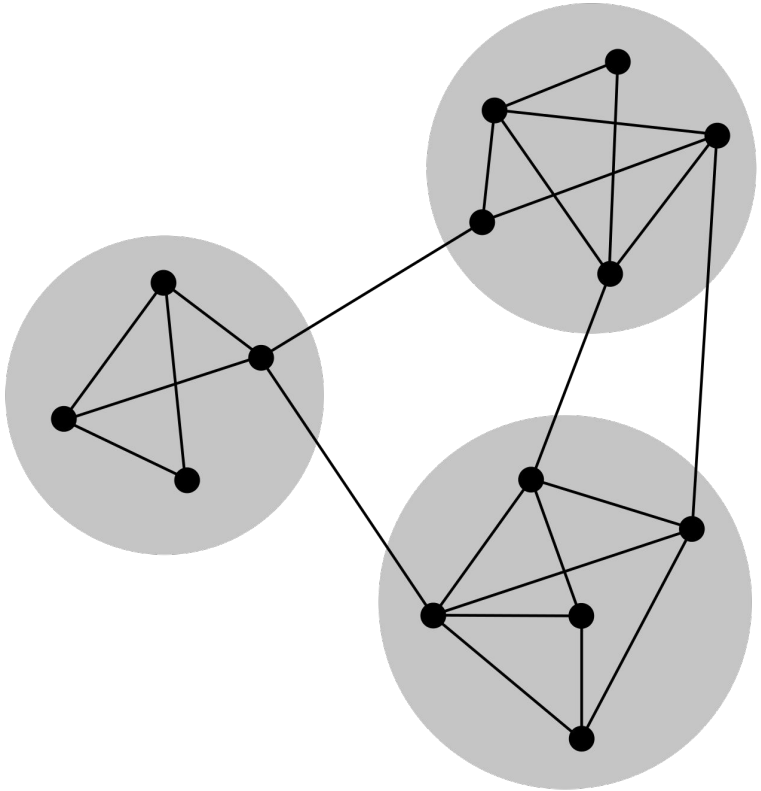


Community structure

A **community** is a locally dense connected subgraph in a network.

Communities are the network equivalent of **clusters**.

As such, there are as many different definitions of communities as there are of clusters.



Traditional clustering

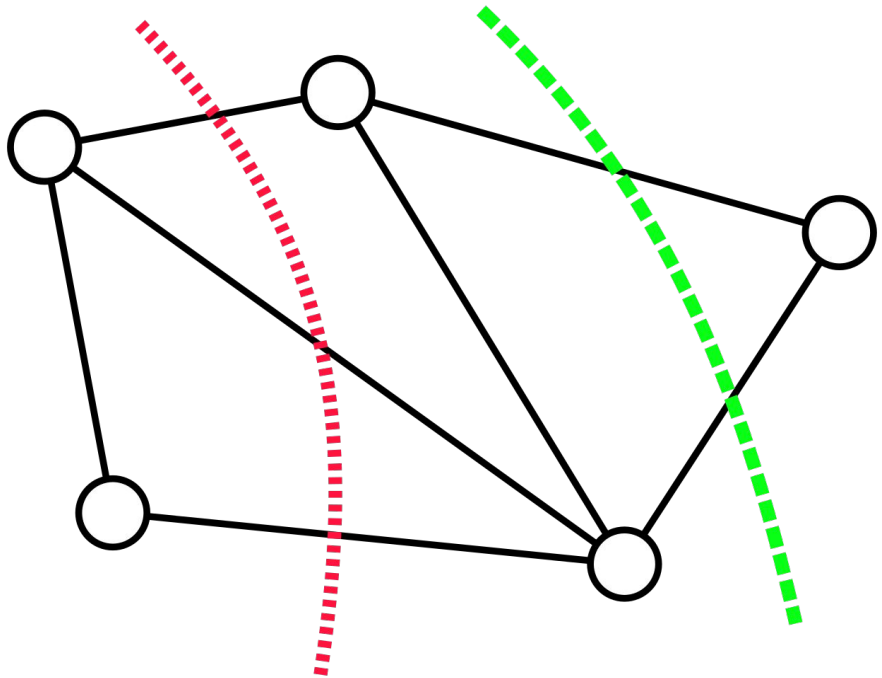
Needs a measure of similarity.

Why not distance? → Embedding problem

Minimum cut

Groups of approximately same size.

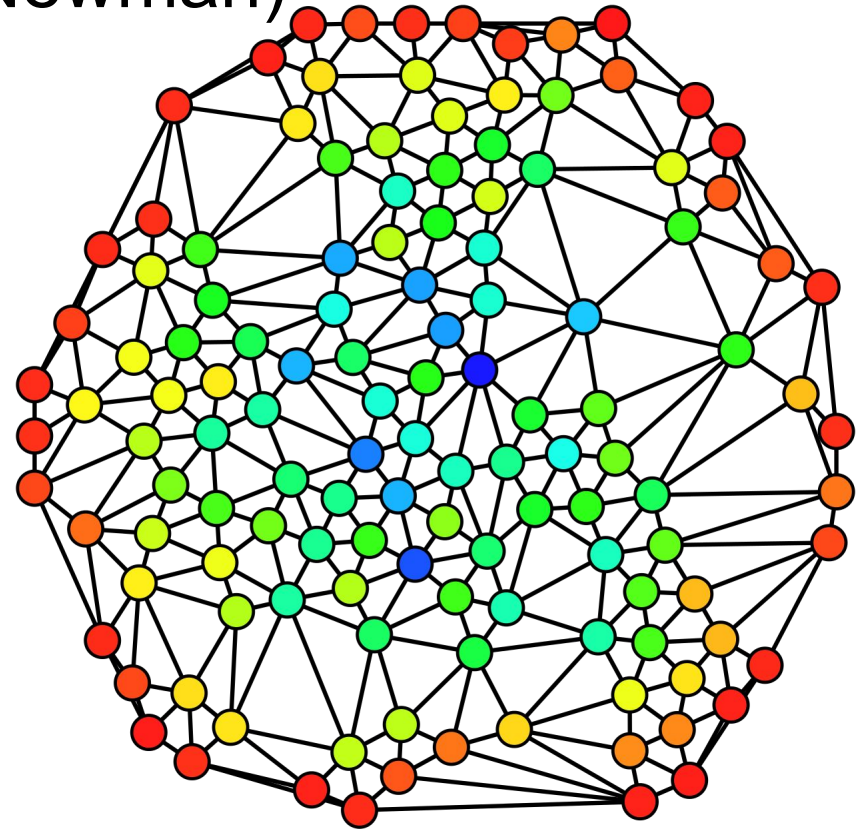
Cuts that minimize the number of edges.



Edge betweenness (Girvan–Newman)

Betweenness is a measure of how many shortest paths pass through a node.

Node with high betweenness separate communities.



Modularity (Louvain)

If networks were connected at random, there would be no communities.

Modularity is the fraction of the edges inside of groups over the expected fraction if edges were distributed at random.

The network partition that gives **maximum modularity** is therefore the one that divides the communities.

Label propagation

Fast algorithm that can be applied in large networks.

It helps to have some of the data labeled first, to narrow down the search.

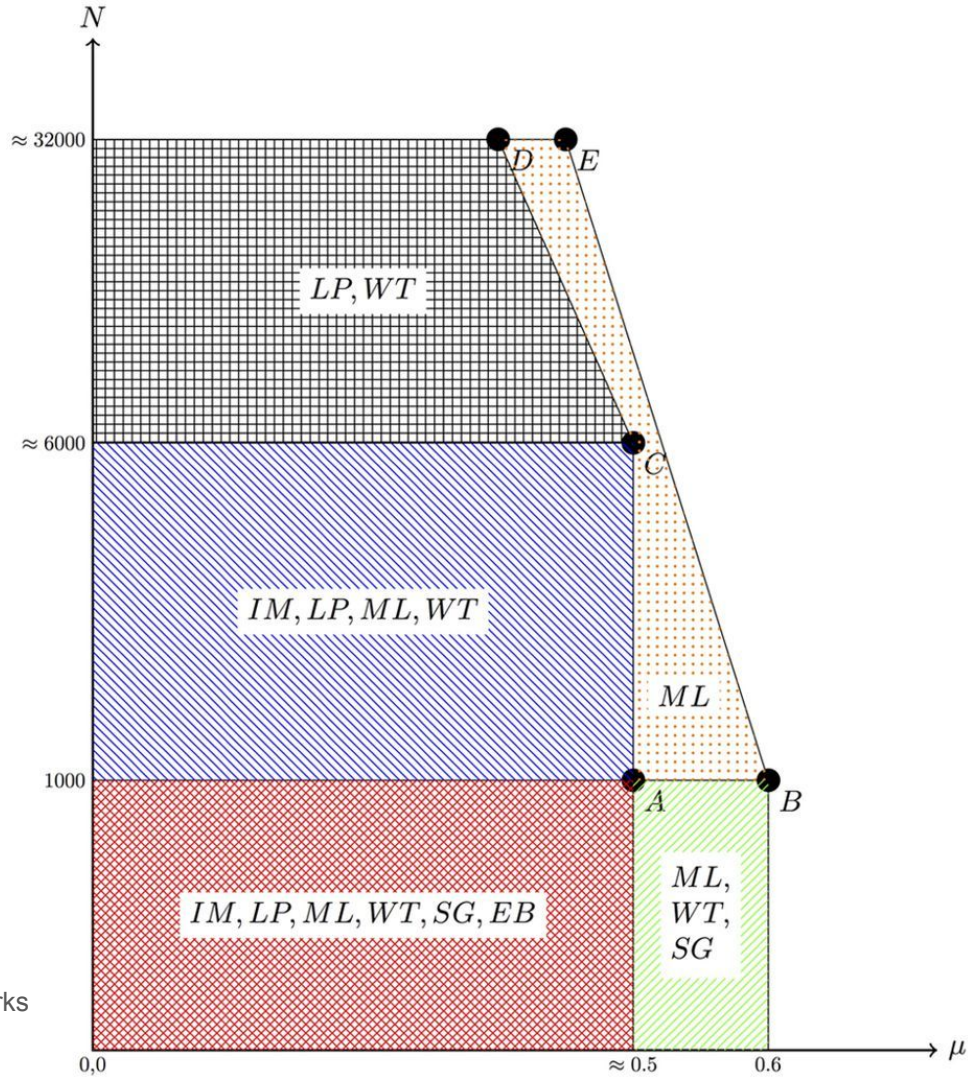
Every node is assigned the label most of it neighbours have.

Community detection

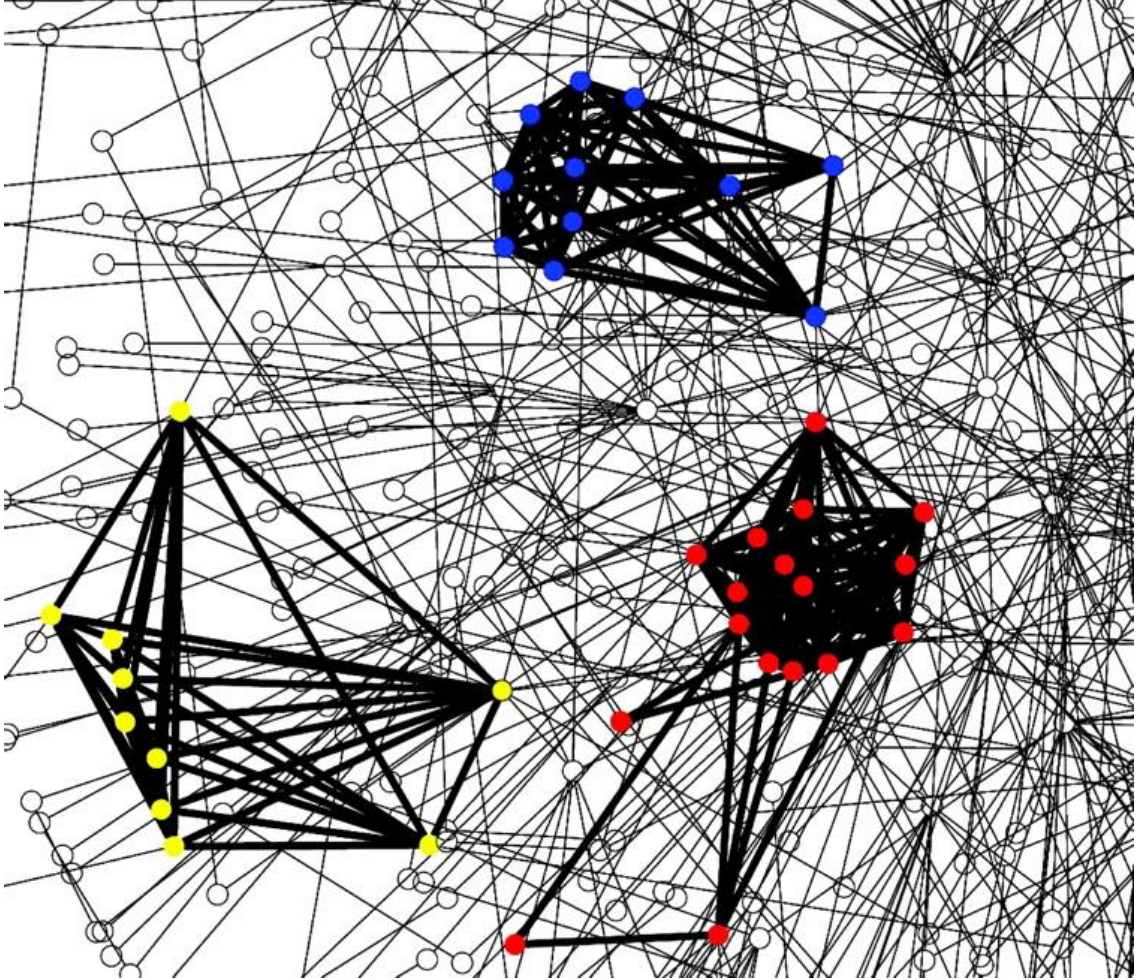
A Comparative Analysis of Community Detection Algorithms on Artificial Networks

Yang, Zhao; Algesheimer, René; Tessone, Claudio J.

Scientific Reports



A Comparative Analysis of Community Detection Algorithms on Artificial Networks
Yang, Zhao; Algesheimer, René; Tessone, Claudio J.
Scientific Reports



Protein complexes and functional modules in molecular networks
Victor Spirin, Leonid A. Mirny
Proceedings of the National Academy of Sciences Oct 2003, 100 (21) 12123-12128; DOI: 10.1073/pnas.2032324100