Graph Theory Basics

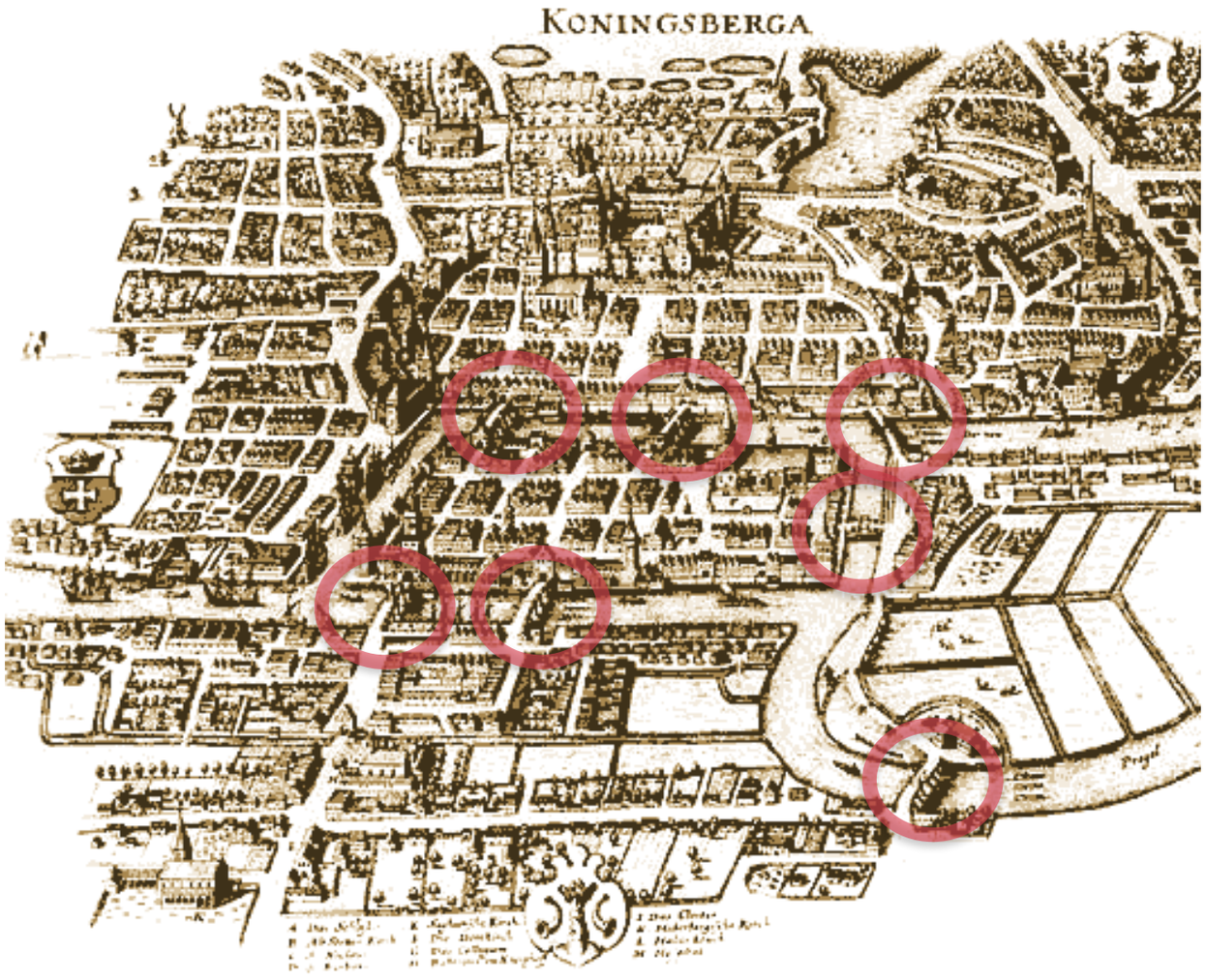
Dr Dorina Thanou

25.02.2025

Outline

- **Networks and graphs**
- Definitions
- Network density
- Pathology
- Connectedness

Back to the XVIIIth century



Konigsberg

Leonhard Euler [1736]

Can one walk across the seven bridges and never cross the same bridge twice?

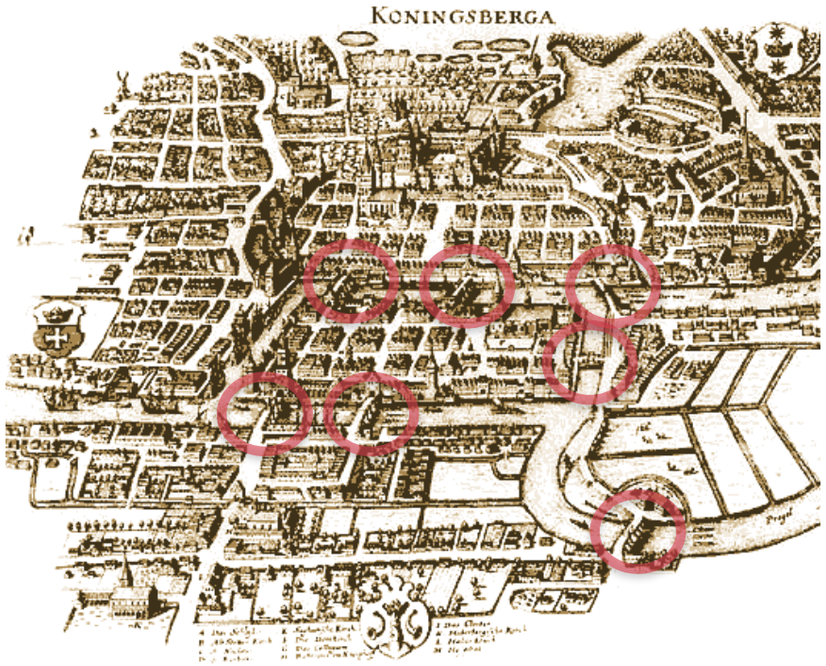
The seven bridges problem

From [1]

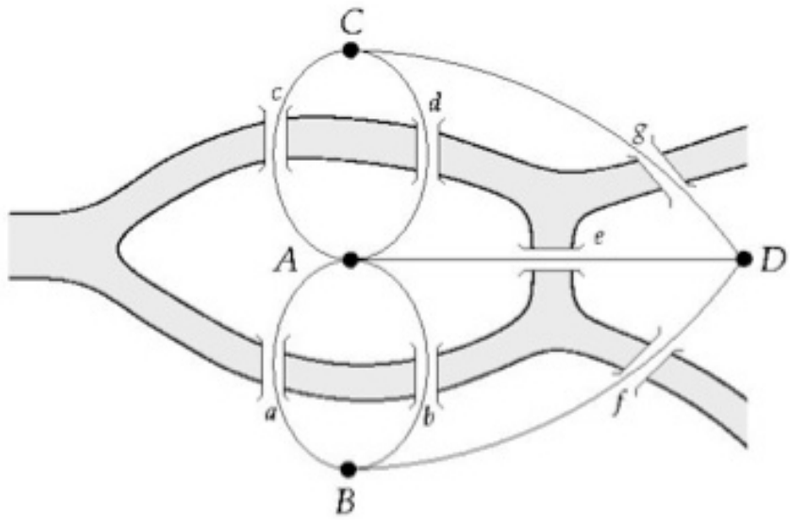
The seven bridges problem

From [1]

Euler's solution



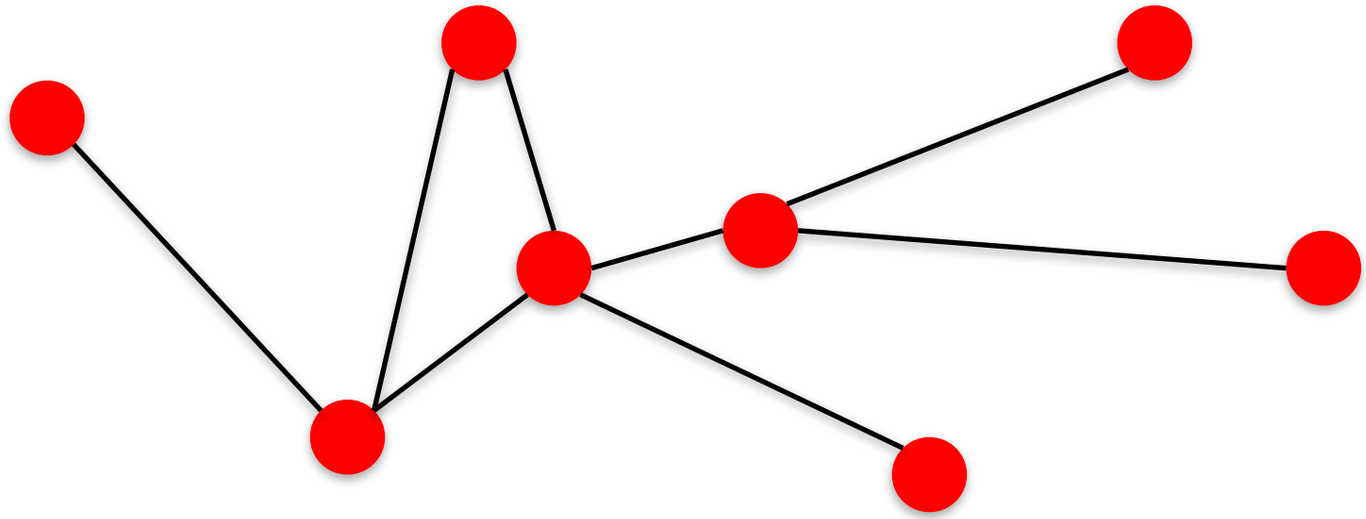
Key representation idea!



- Euler's theorem:
  - If a graph has more than two nodes of odd degree, there is no path
  - If a graph is connected and has no odd degree nodes, it has at least one path
- Conclusion: Networks have properties encoded in their structure

Graph theory offers many analysis tools to use networks for various applications: from clustering to classification, visualization, recommendation

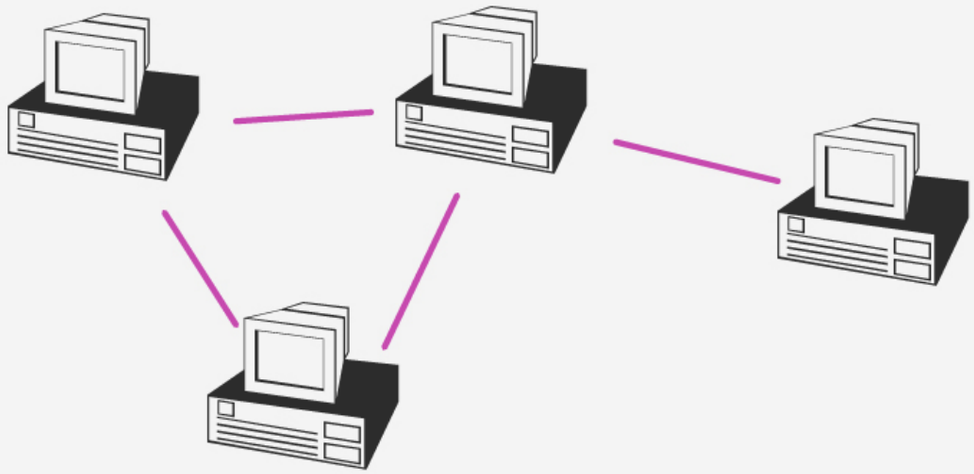
Networks and graphs

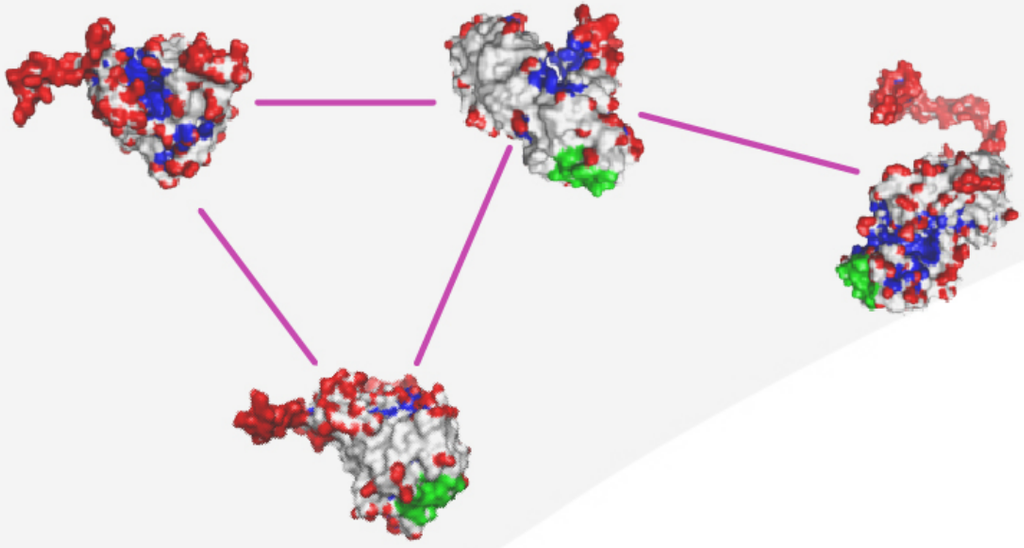


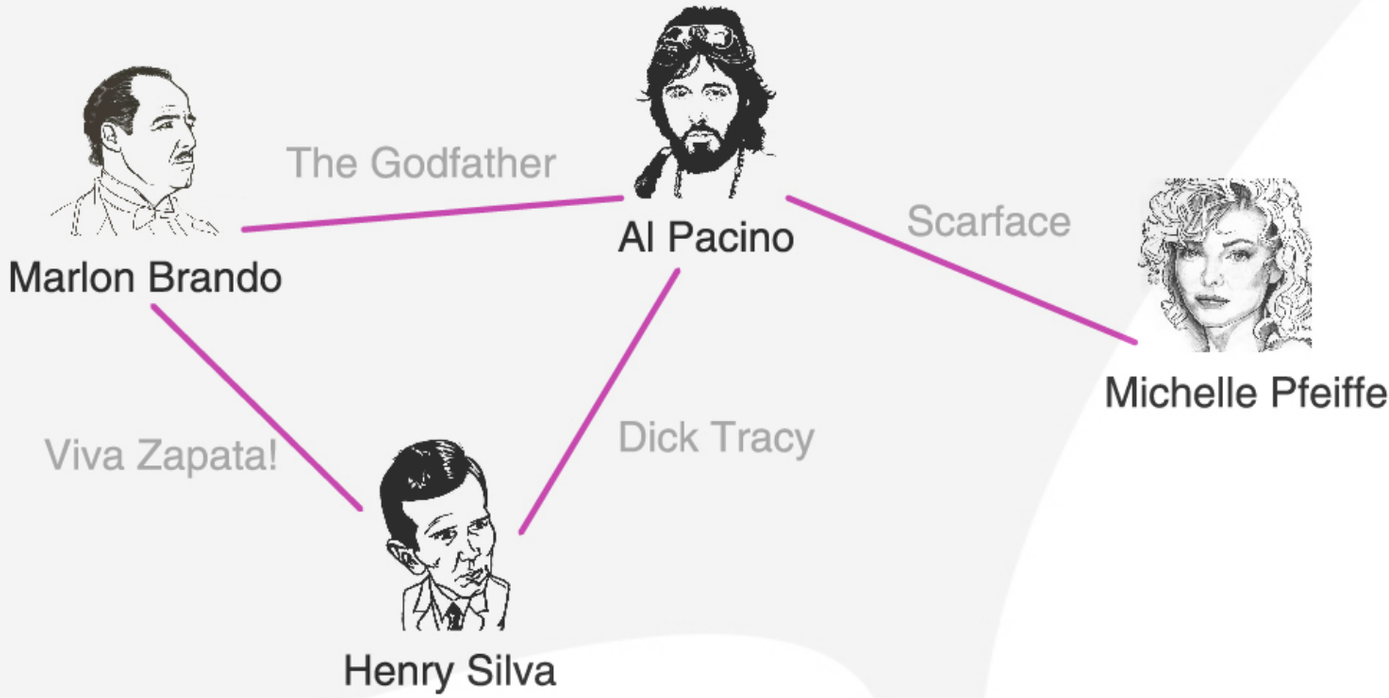
- Elements of a complex system
  - Vertex set $\mathcal{V}$ of $N = |\mathcal{V}|$ nodes represent the components of the system
  - Edge set $\mathcal{E}$ of $M = |\mathcal{E}|$ links capture the interactions between components
  - Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ describes the system

A **network** generally refers to the real system, a **graph** is its mathematical description

Graphs provide generic description

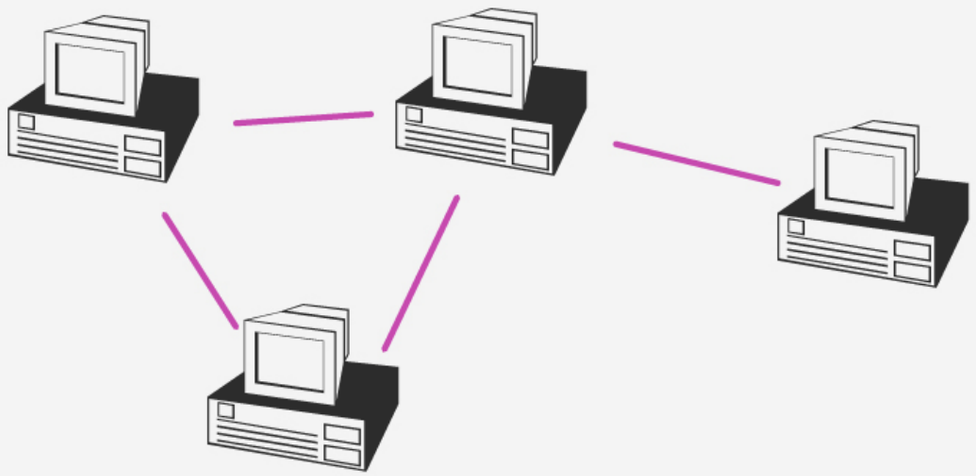


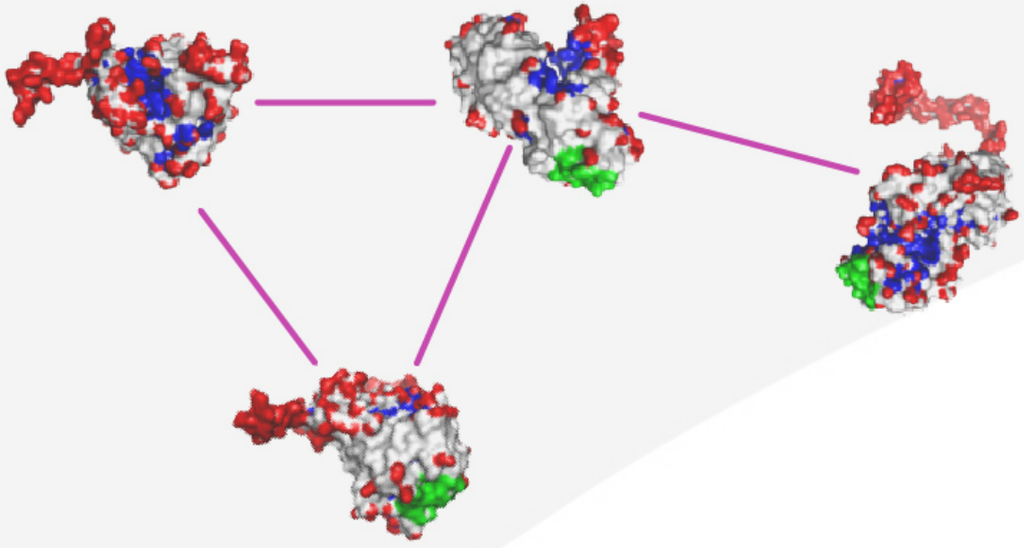


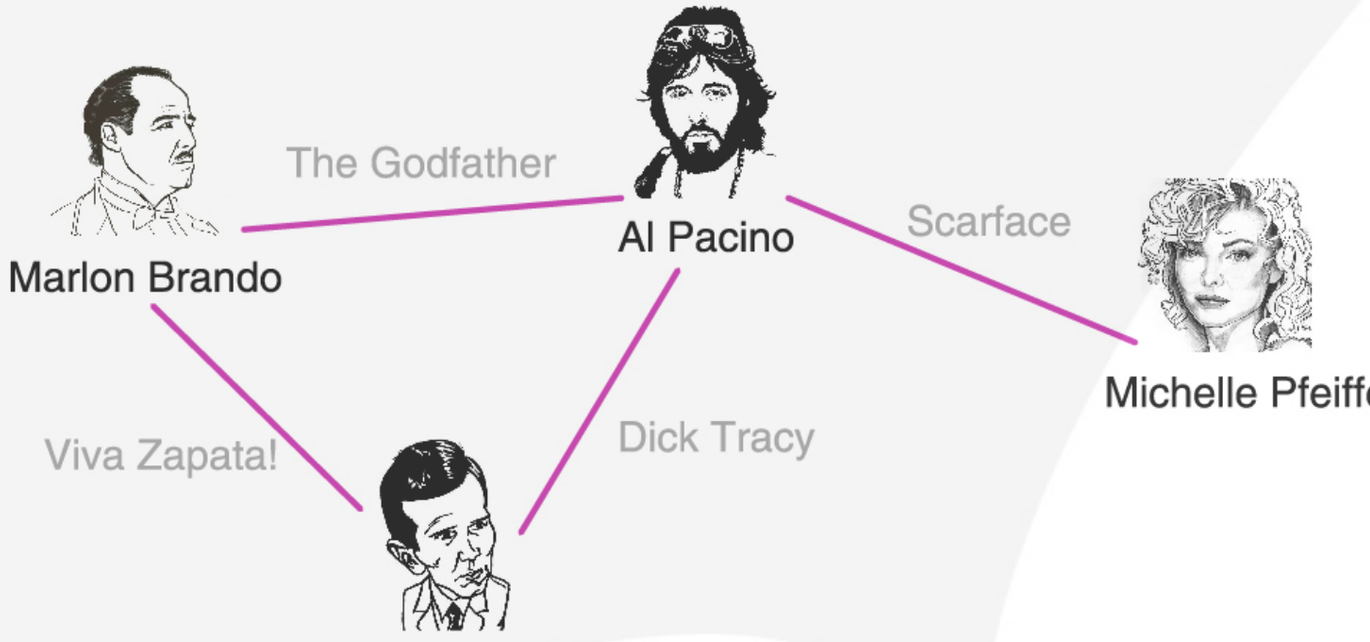


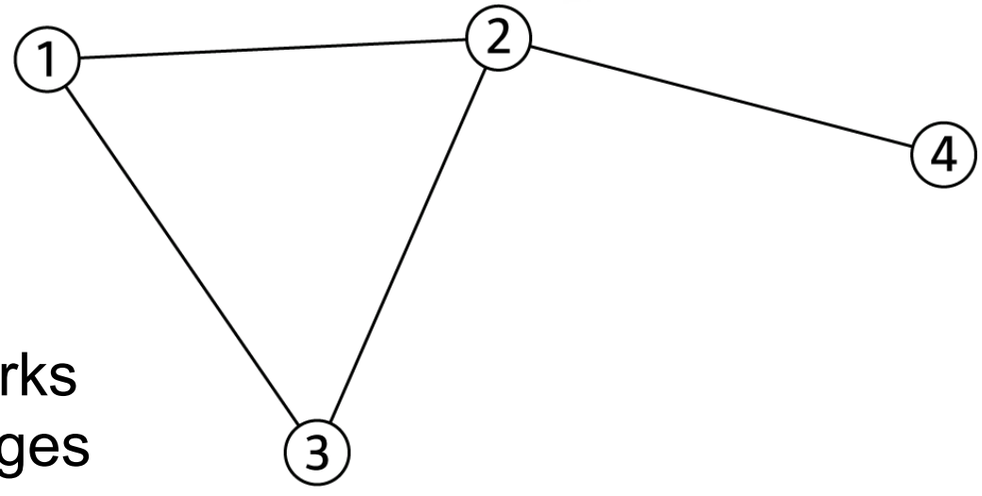
A common language to study systems: All networks are represented by a graph with 4 vertices and 4 edges

Graphs provide generic description







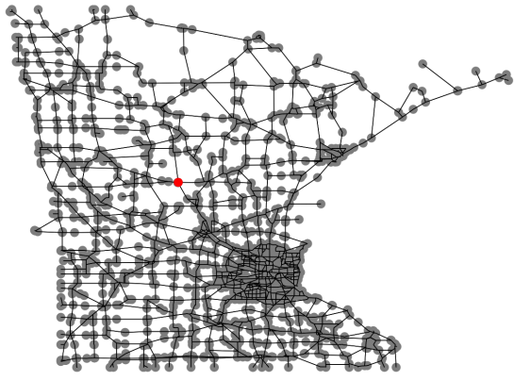


A common language to study systems: All networks are represented by a graph with 4 vertices and 4 edges

Importance of the representation

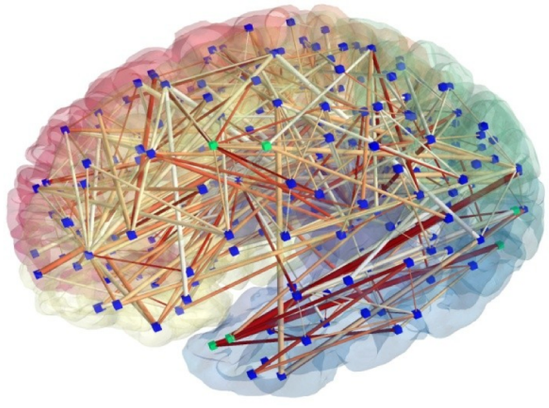
- The way we design the network will determine the nature of questions we can study

Natural graphs

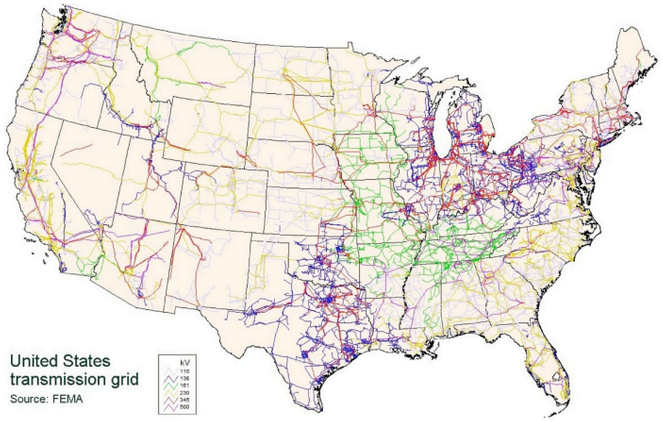


Minnesota Road Network

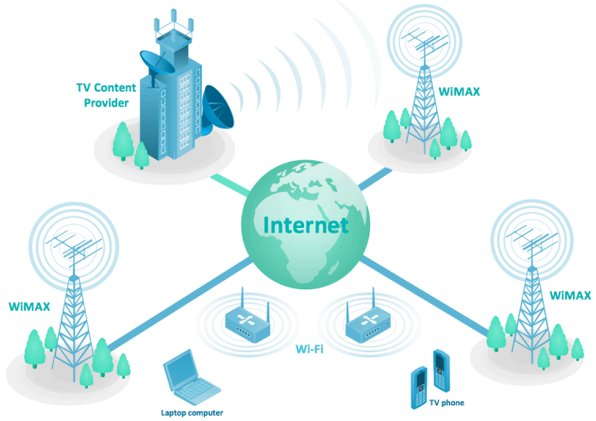
Facebook



Brain Connectivity



US Electrical Network



Telecommunication Network

- Natural networks (i.e., not artificially constructed) can be directly represented by graphs

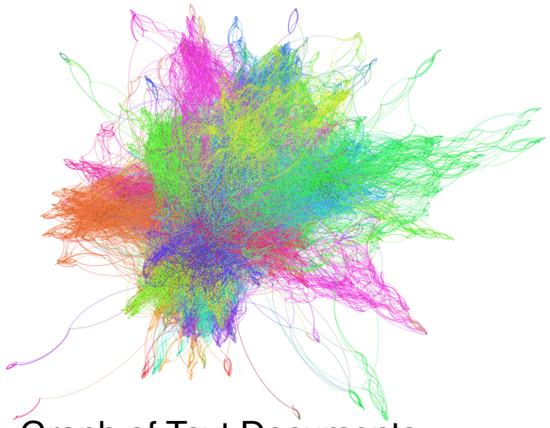
Common natural networks

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	Number of nodes	Number of links
Internet	Routers	Internet connections	Undirected	192,244	609,066
WWW	Webpages	Links	Directed	325,729	1,497,134
Power Grid	Power plants, transformers	Cables	Undirected	4,941	6,594
Mobile Phone Calls	Subscribers	Calls	Directed	36,595	91,826
Email	Email addresses	Emails	Directed	57,194	103,731
Science Collaboration	Scientists	Co-authorship	Undirected	23,133	93,439
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908
Citation Network	Paper	Citations	Directed	449,673	4,689,479
E. Coli Metabolism	Metabolites	Chemical reactions	Directed	1,039	5,802
Protein Interactions	Proteins	Binding interactions	Undirected	2,018	2,930

From [1]

Graphs constructed from data structure

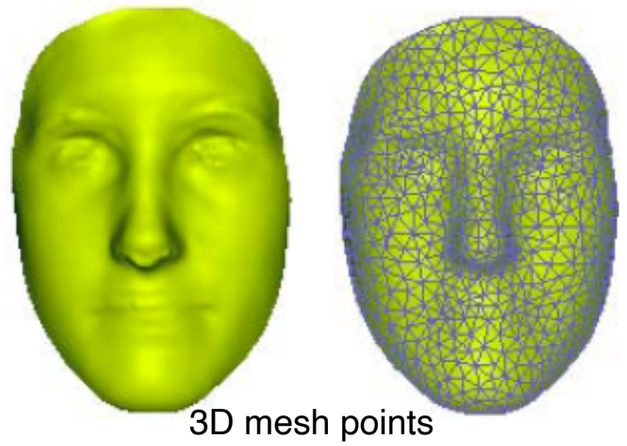
- Graphs provide a way to identify hidden structure in high-dimensional data (e.g., through manifold learning)



Graph of Text Documents
20newsgroups

MNIST Image Graph

GTZAN Music Graph

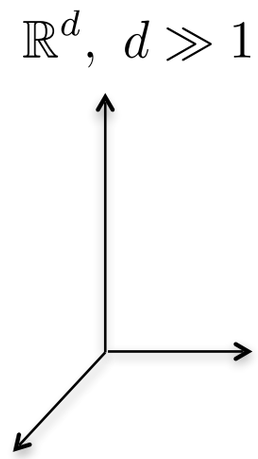


3D mesh points

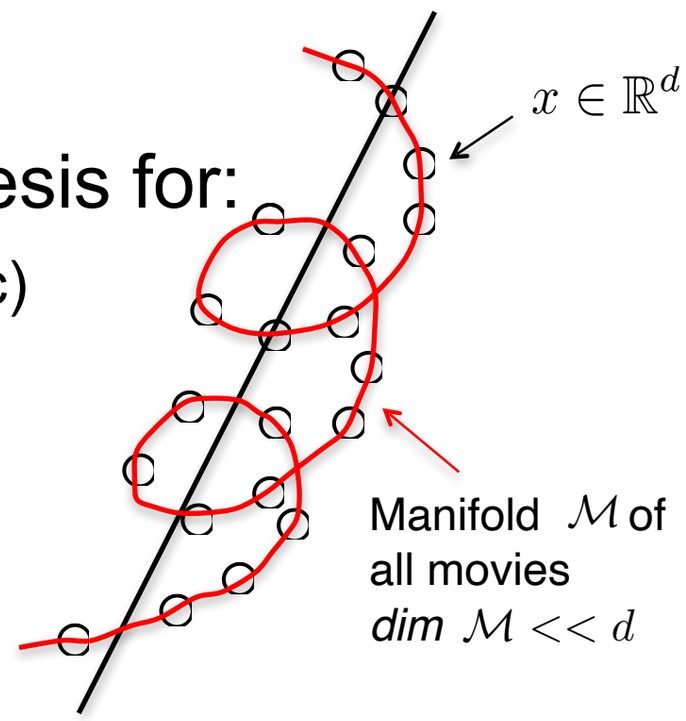
- They can be constructed from data, using modeling assumptions or domain expertise

Data structure and manifolds

- **Manifold assumption:** High-dim data are sampled from a low-dim manifold
  - Let *x* be a movie, each movie is defined by *d* features/attributes like genre, actors, release year, origin country, etc such that *x* in R^d
  - Then we can make the assumption that all movies form a manifold in R^d



- Assumption validity: It is a good working hypothesis for:
  - Several types of data (images, text documents, music, etc)
  - Most data science tasks (classification, visualization, recommendation, etc)



From manifolds to graphs

- Graphs can be seen as manifold sampling: the manifold information is encoded by neighborhood graphs

[Belkin]

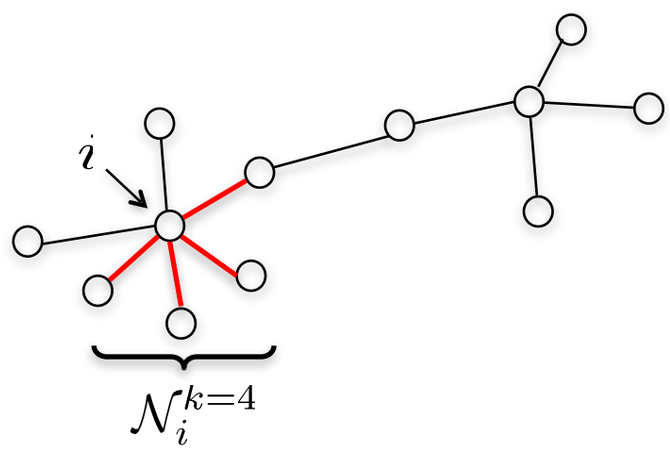
- Neighborhood Graphs: k-NN graphs are among the most popular

$$W_{ij} = \begin{cases} e^{-\frac{\mathrm{dist}(x_i,x_j)^2}{\sigma^2}} & \text{if } j \in \mathcal{N}_i^k \\ 0 \text{ otherwise} \end{cases}$$

distance between samples

neighbourhood

scale parameter



How to construct graphs from data?

- No universal recipe is available
- It often depends on data and the specific analytical task
- Domain expertise and good practices are key

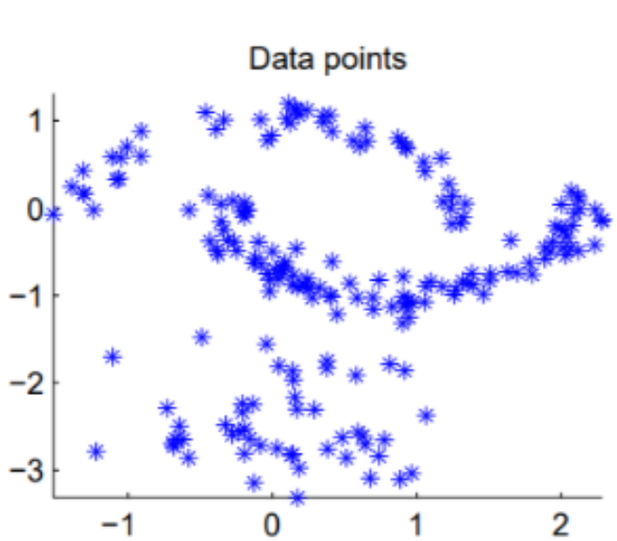
- Design choices:
  - Which type of graph (e.g., fully connected, sparse)?
  - Which distance function to use between data points (e.g., Euclidean distance, cosine similarity, KL distance)?
  - Which data features are more meaningful to compare (e.g., raw features, hand-crafted features, learned features)?

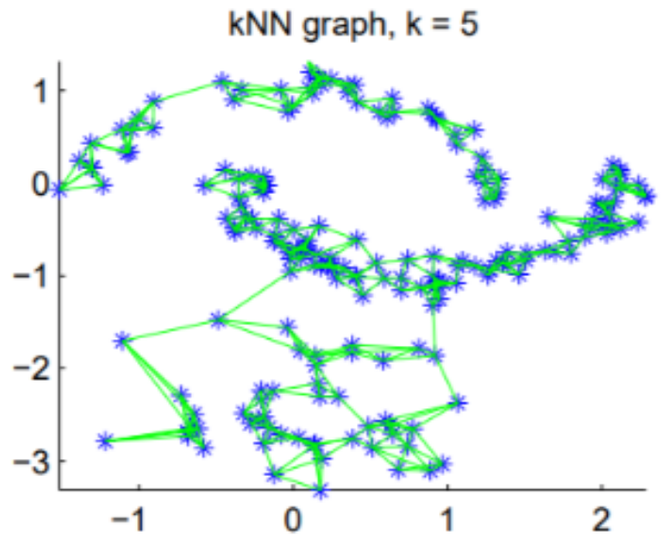
Illustrative example

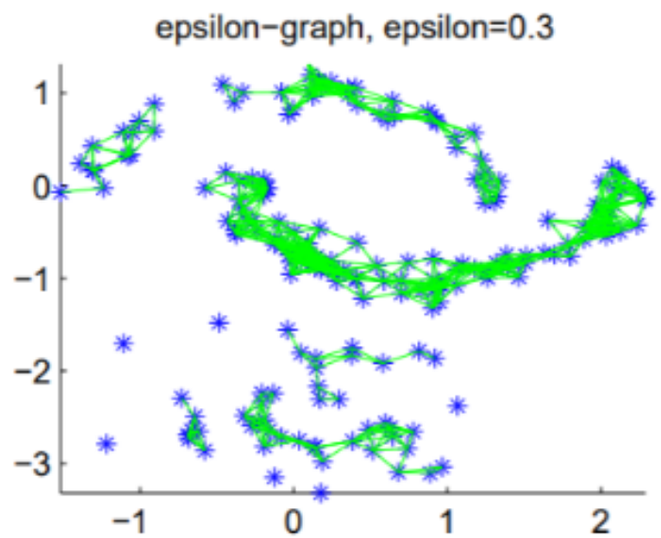
- The graph is usually constructed based on some feature/data similarity:
  - Compute a graph kernel matrix (i.e., fully connected graph); usually an RBF kernel

$$K(i,j) = \exp\left(-\frac{\|x_i - x_j\|_2^2}{2\sigma^2}\right)$$

  - Sparsify the kernel matrix to obtain a connectivity matrix
    - ϵ -nearest neighbor graph
    - K-nearest neighbor graph



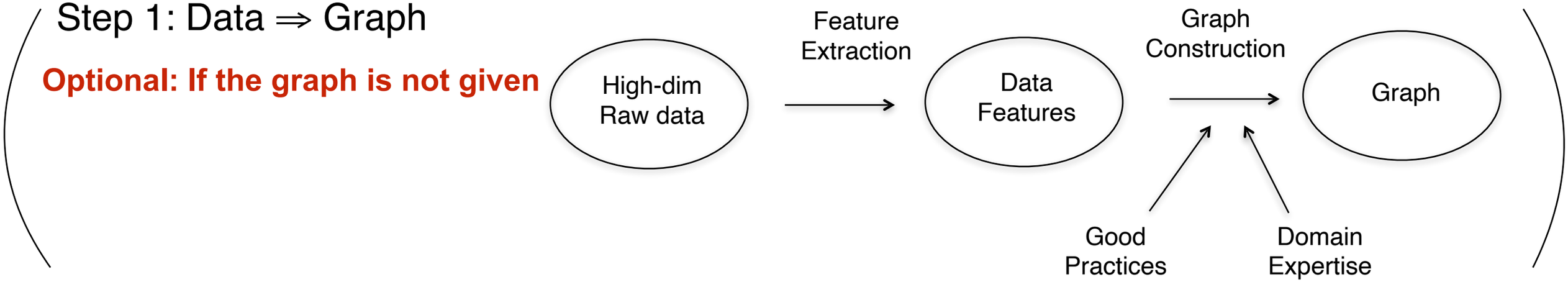




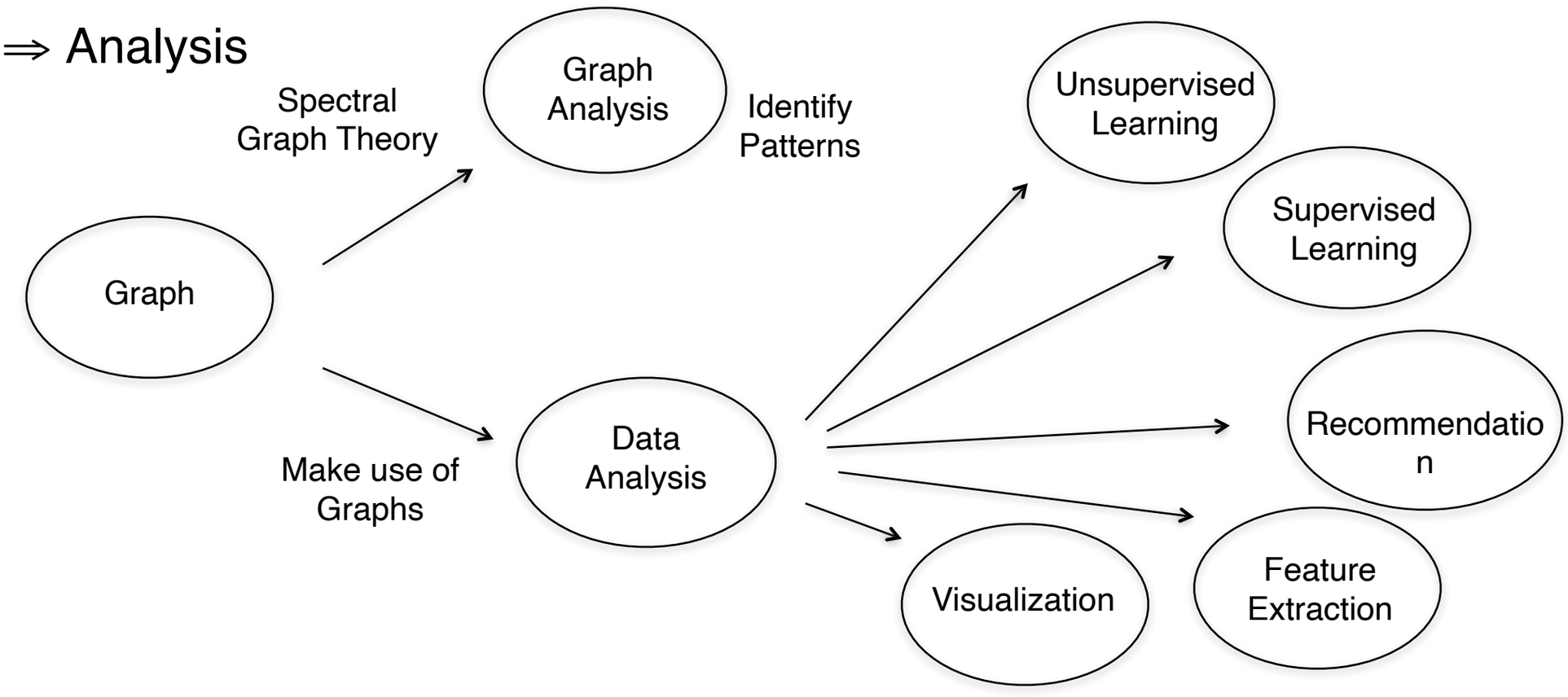
Structured data analysis pipeline

Step 1: Data ⇒ Graph

Optional: If the graph is not given



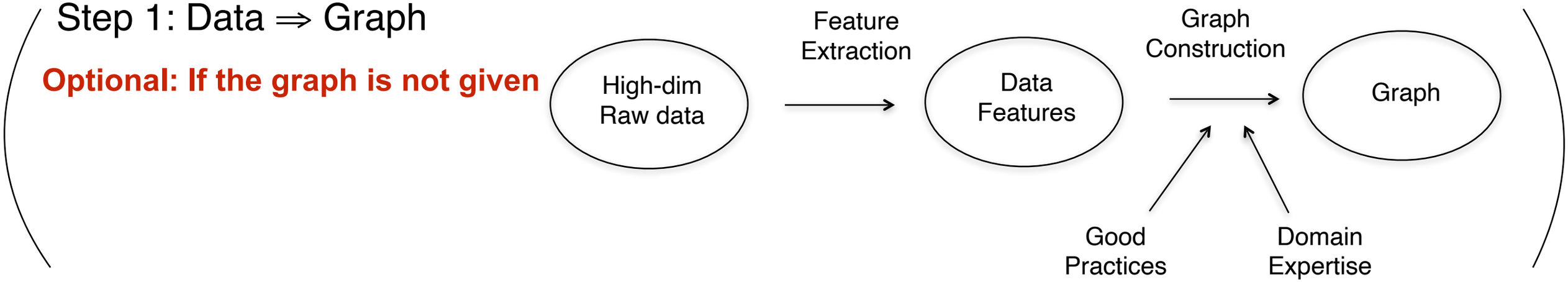
Step 2: Graph ⇒ Analysis



Structured data analysis pipeline

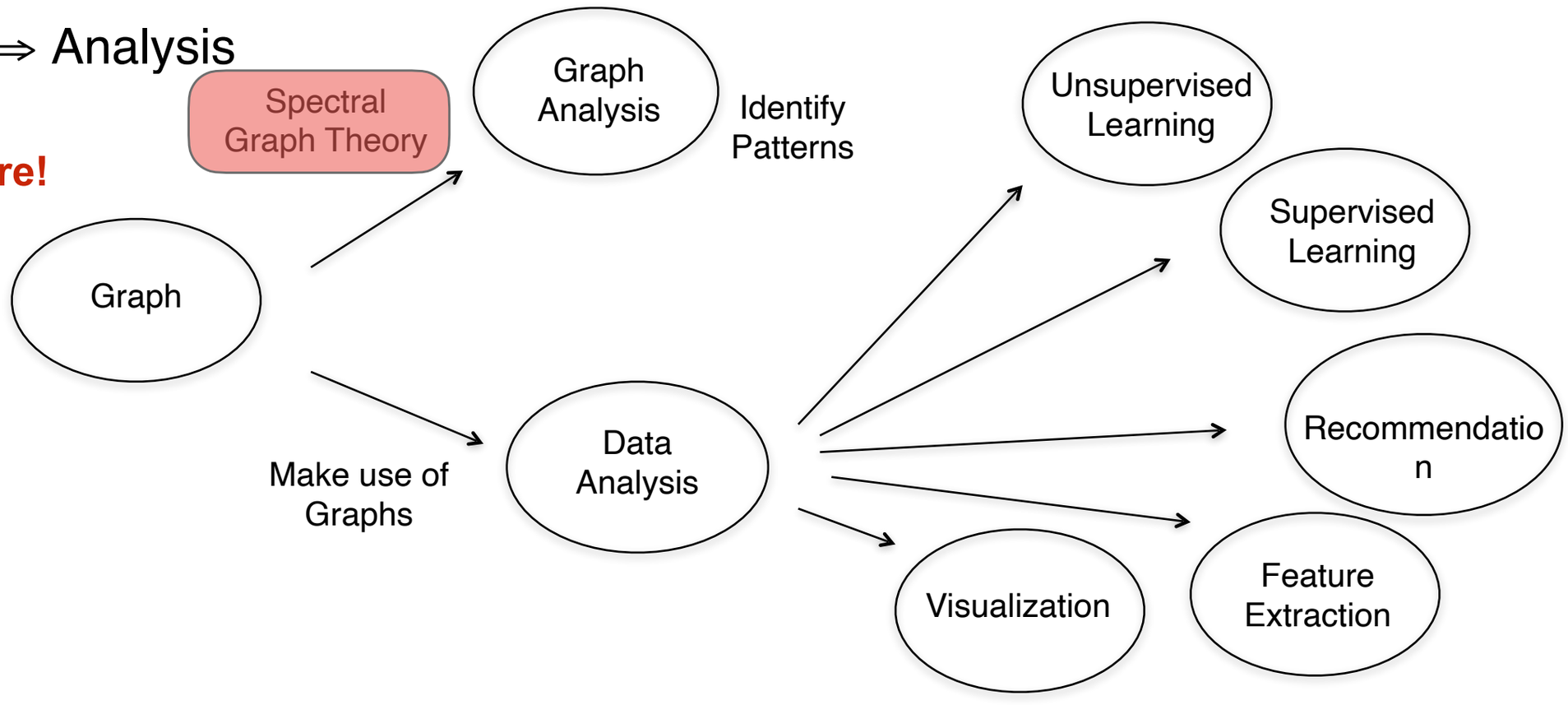
Step 1: Data ⇒ Graph

Optional: If the graph is not given



Step 2: Graph ⇒ Analysis

Following lecture!

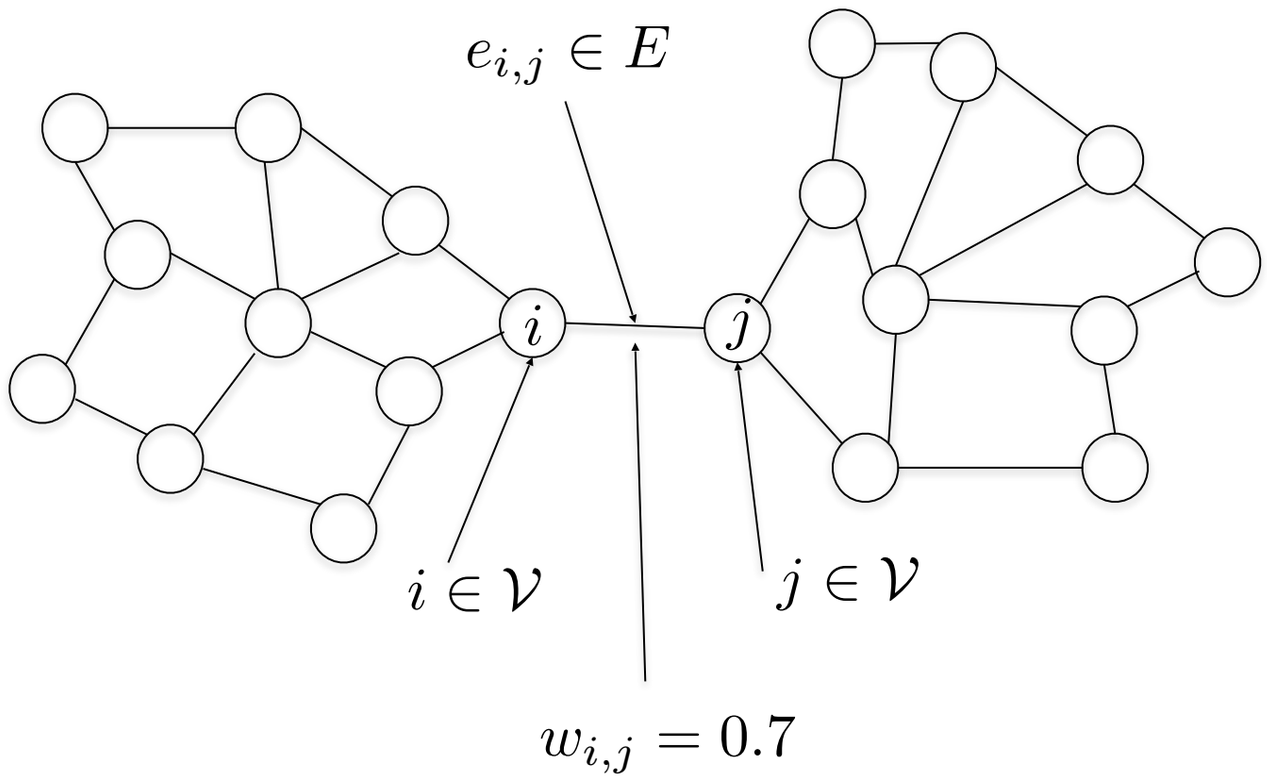


Overview

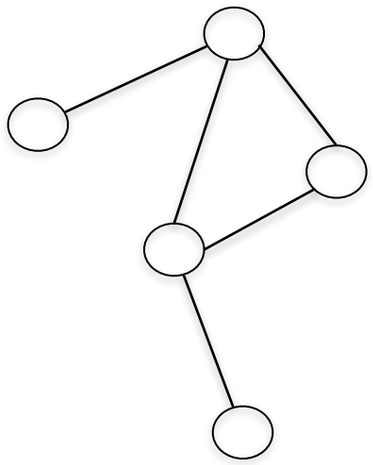
- Networks and graphs
- **Definitions**
- Network density
- Pathology
- Connectedness

Weighted graphs

- A **weighted** graph is fully defined by $\mathcal{G} = (\mathcal{V}, \mathcal{E}, W)$
  - $\mathcal{V}$ is the set of vertices, $\mathcal{E}$ the set of edges, and W the weight or similarity matrix
  - The total number of vertices is $|\mathcal{V}| = N$ and the total number of edges is $|\mathcal{E}| = M$
  - The graph is **unweighted** if the edge weights are binary, i.e., $e_{i,j} \in \{0, 1\}, \forall i, \forall j$

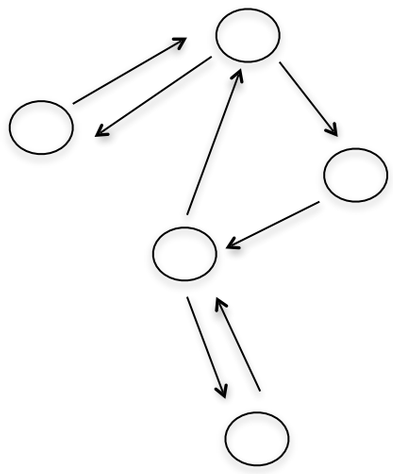


Directed graphs



undirected

- Links: undirected (symmetrical)
  - coauthorship links
  - actor network
  - protein interactions



directed

- Links: directed (arcs)
  - URLs on the www
  - phone calls
  - metabolic reactions
- Directed graph = Digraph

Super-fast statistics recap…

For samples of *N* values:

Average (mean)

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

nth moment

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

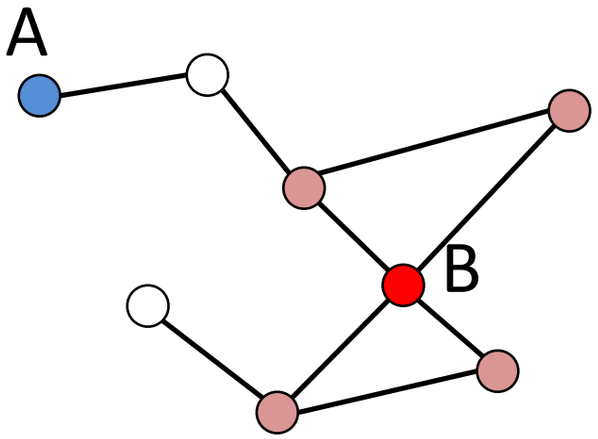
Standard deviation

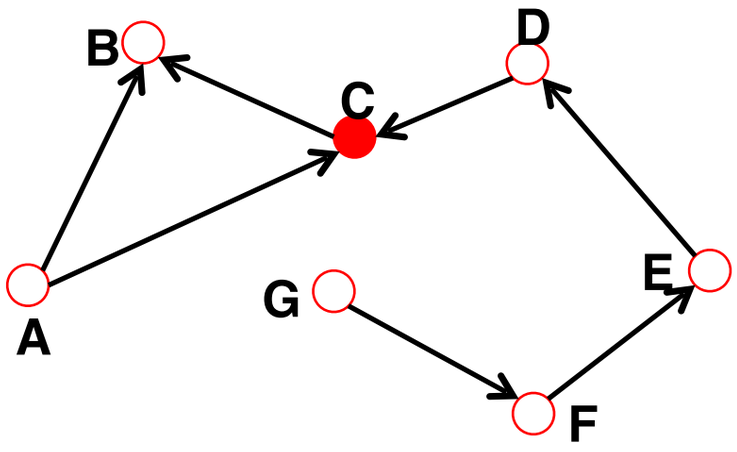
$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(x_i - \langle x \rangle\right)^2}$$

Distribution of x

$$p_x = \frac{1}{N}\sum_{i} \delta_{x,x_i} \qquad \text{with} \qquad \sum_x p_x = 1 \left(\int p_x dx = 1 \right)$$

Node degree





- Undirected graph
  - degree D_i = number of connected edges

$$D_A = 1 \qquad D_B = 4 \qquad M = \frac{1}{2}\sum_{i=1}^{N} D_i$$

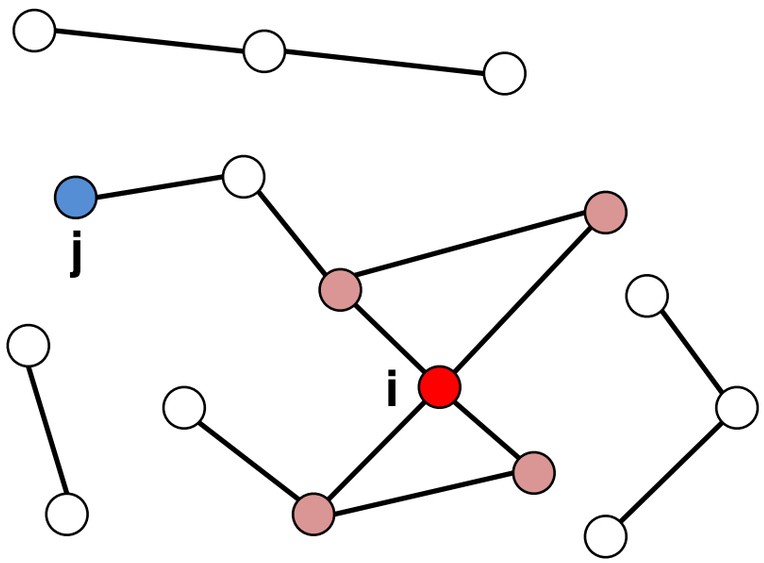
- Directed graph
  - in-degree D_i^{in} = sum of incident edges
  - out-degree D_i^{out} = sum of outgoing edges
  - (total) degree $D_i = D_i^{in} + D_i^{out}$
  - source = node with $D_i^{in} = 0$
  - sink = node with $D_i^{out} = 0$

$$D_C^{in} = 2 \qquad D_C^{out} = 1 \qquad D_C = 3$$

- For weighted graphs, the degree is the sum of the edge weights, and not the number of edges, i.e., $D_i = \sum_{j \in \mathcal{V}} W_{ij}$

Average Degree

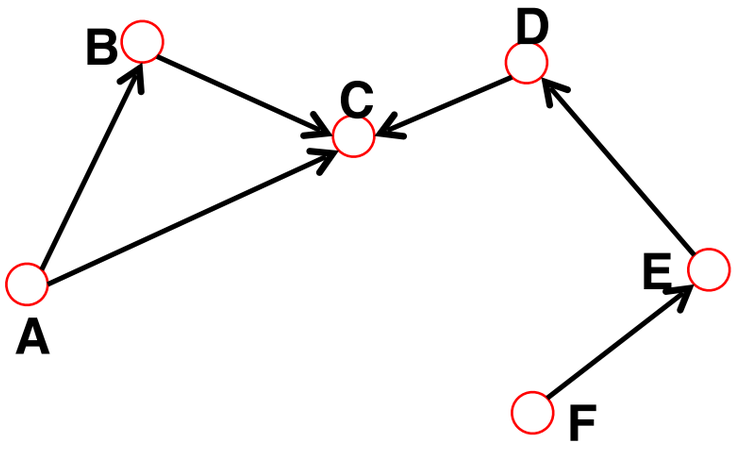
Undirected graphs



Average degree: $\langle D \rangle = \frac{1}{N} \sum_{i=1}^{N} D_i$

$\langle D \rangle = \frac{2M}{N}$

Directed graphs



Average degrees: $\langle D^{in} \rangle = \frac{1}{N} \sum_{i=1}^{N} D_i^{in}$

$\langle D^{out} \rangle = \frac{1}{N} \sum_{i=1}^{N} D_i^{out}$

$\langle D^{in} \rangle = \langle D^{out} \rangle = \frac{M}{N}$

Average degree in reference networks

Network type	Nodes	Links	Directed/Undirected	N	M	$\langle D \rangle$
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

From [1]

Degree distribution

- p_k : probability that a node i has degree $D_i = k$

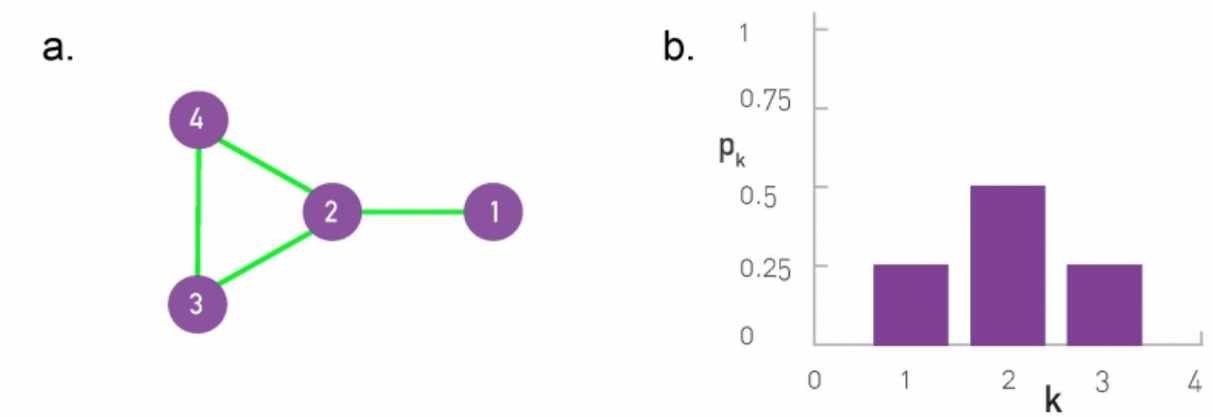
$$p_k = \frac{N_k}{N} \qquad \langle D \rangle = \sum_{k}^{\infty} k p_k$$

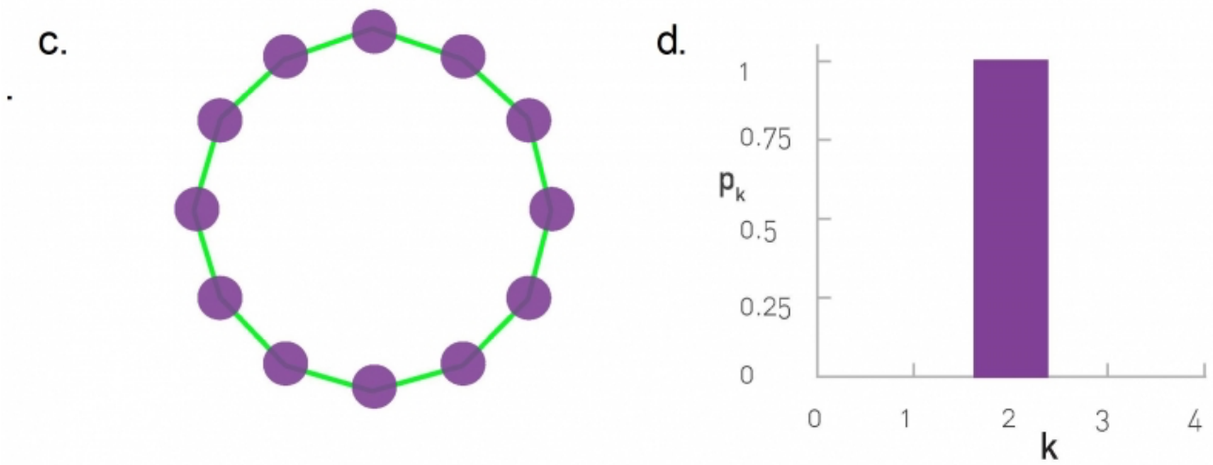
  - with N_k the number of nodes of degree k

- Normalization conditions:

$$\sum_{k=0}^{\infty} p_k = 1$$

- Continuous' version : $p(k)$
  - with probability that a node has degree between k_1 and k_2 given by $\int_{k_1}^{k_2} p(k)dk$, and $\int_{0}^{\infty} p(k)dk = 1$





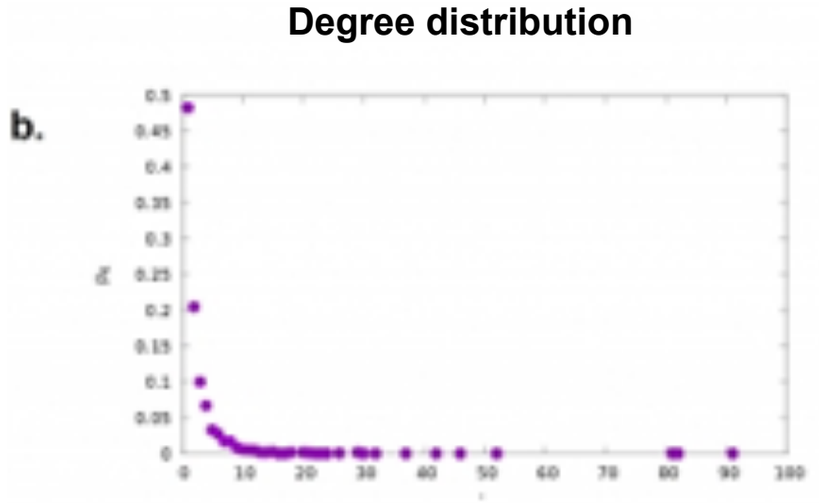
From [1]

Degree distribution in a real network

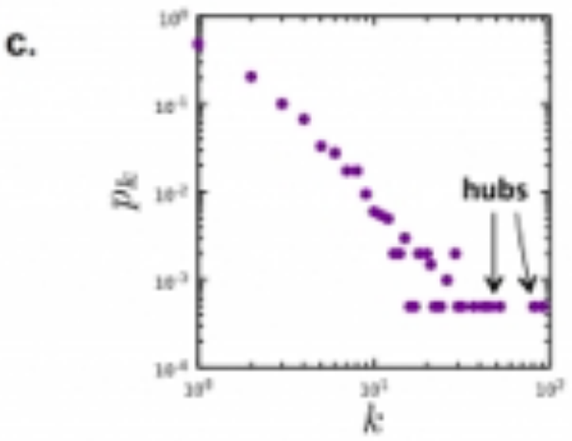
Protein interaction network of yeast

From [1]

Degree distribution



Degree distribution (log-log)



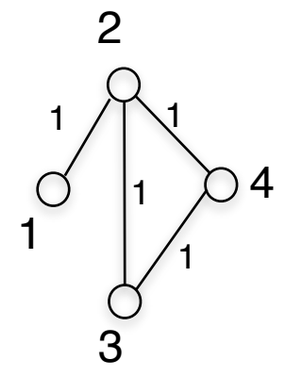
Adjacency matrix

- The **adjacency matrix** A (aka **similarity**, or **weight matrix** W) captures the connectivity pattern of the network

- Binary matrix (unweighted graph): $w_{i,j} \in \{0, 1\}$

$$W_{i,j} = \begin{cases} 1, \text{ if } (i,j) \in E \\ 0, \text{ otherwise} \end{cases}$$

G=

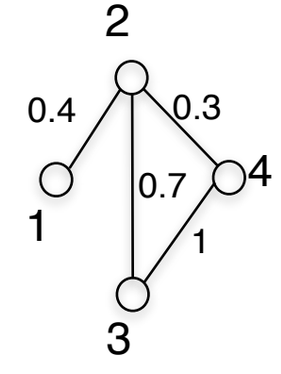


$$\Rightarrow W = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

- Weighted matrix: $w_{i,j} \in [0, 1]$ (commonly normalized)

$$W_{i,j} = \begin{cases} w_{i,j} \in [0,1], \text{ if } (i,j) \in E \\ 0, \text{ otherwise} \end{cases}$$

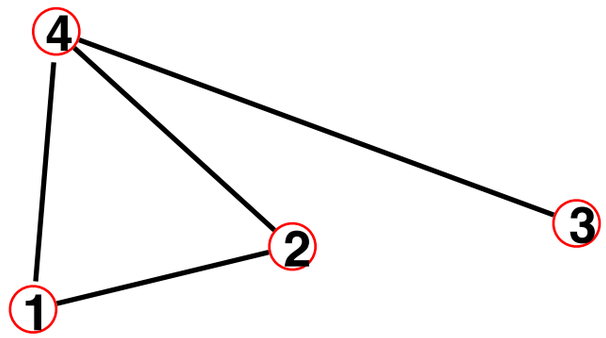
G=



$$\Rightarrow W = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 0.4 & 0 & 0 \\ 0.4 & 0 & 0.7 & 0.3 \\ 0 & 0.7 & 0 & 1 \\ 0 & 0.3 & 1 & 0 \end{bmatrix} \end{array}$$

Adjacency matrix and degrees

Undirected graphs



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

$$W_{ij} = W_{ji}$$

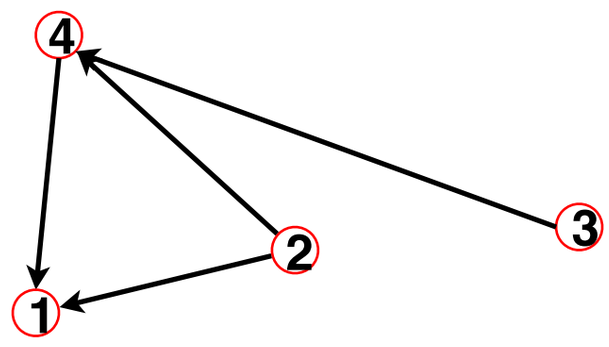
$$W_{ii} = 0$$

$$D_i = \sum_{j=1}^{N} W_{ij}$$

$$D_j = \sum_{i=1}^{N} W_{ij}$$

$$M = \frac{1}{2}\sum_{i=1}^{N} D_i = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}$$

Directed graphs



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

$$W_{ij} \neq W_{ji}$$

$$W_{ii} = 0$$

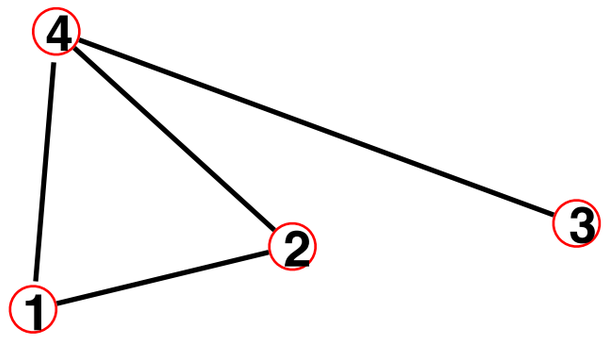
$$D_i^{out} = \sum_{j=1}^{N} W_{ij}$$

$$D_j^{in} = \sum_{i=1}^{N} W_{ij}$$

$$M = \sum_{i=1}^{N} D_i^{in} = \sum_{j=1}^{N} D_j^{out} = \sum_{i,j=1}^{N} W_{ij}$$

Adjacency matrix and degrees

Undirected graphs



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

$$W_{ij} = W_{ji}$$

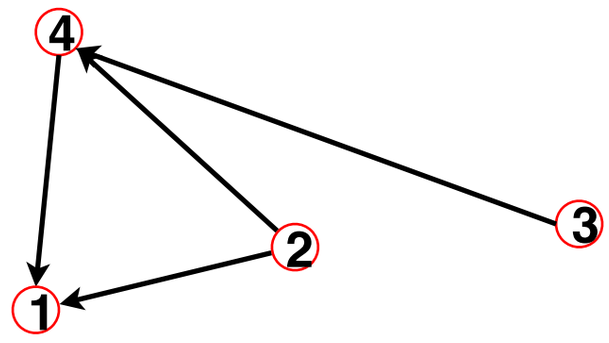
$$W_{ii} = 0$$

$$D_i = \sum_{j=1}^{N} W_{ij}$$

$$D_j = \sum_{i=1}^{N} W_{ij}$$

$$M = \frac{1}{2}\sum_{i=1}^{N} D_i = \frac{1}{2}\sum_{i,j=1}^{N} W_{ij}$$

Directed graphs



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

$$W_{ij} \neq W_{ji}$$

$$W_{ii} = 0$$

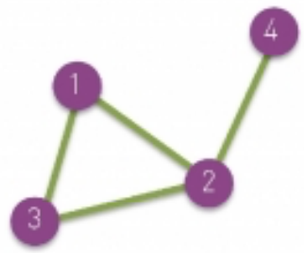
$$D_i^{out} = \sum_{j=1}^{N} W_{ij}$$

$$D_j^{in} = \sum_{i=1}^{N} W_{ij}$$

$$M = \sum_{i=1}^{N} D_i^{in} = \sum_{j=1}^{N} D_j^{out} = \sum_{i,j=1}^{N} W_{ij}$$

Graphology

a. Undirected

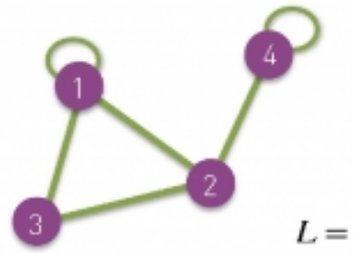


$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

$$M = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \langle D \rangle = \frac{2M}{N}$$

b. Self-loops



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

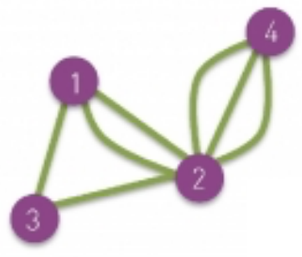
$\exists i,\ A_{ii} \neq 0 \qquad A_{ij} = A_{ji}$

$$L = \frac{1}{2}\sum_{i,j=1, i\neq j}^{N} A_{ij} + \sum_{i=1}^{N} A_{ii} \qquad ?$$

Internet, power grid, science collaboration networks

Protein interaction network, www

c. Multigraph
(undirected)

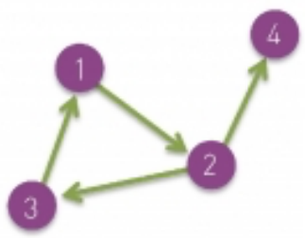


$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

$$M = \frac{1}{2}\sum_{i,j=1}^{N} A_{ij} \qquad \langle D \rangle = \frac{2M}{N}$$

d. Directed



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

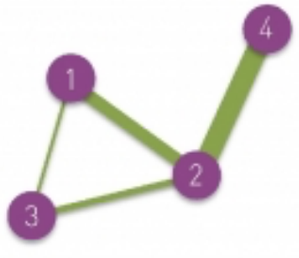
$A_{ij} \neq A_{ji}$

$$M = \sum_{i,j=1}^{N} A_{ij} \qquad \langle D \rangle = \frac{M}{N}$$

Social networks, collaboration networks

WWW, mobile phone calls, citation network

e. Weighted
(undirected)

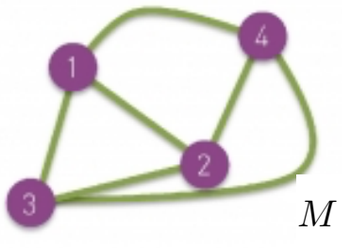


$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

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

$$\langle D \rangle = \frac{2M}{N}$$

f. Complete Graph
(undirected)



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

$A_{ii} = 0 \qquad A_{i\neq j} = 1$

$$M = M_{max} = \frac{N(N-1)}{2} \qquad \langle D \rangle = N - 1$$

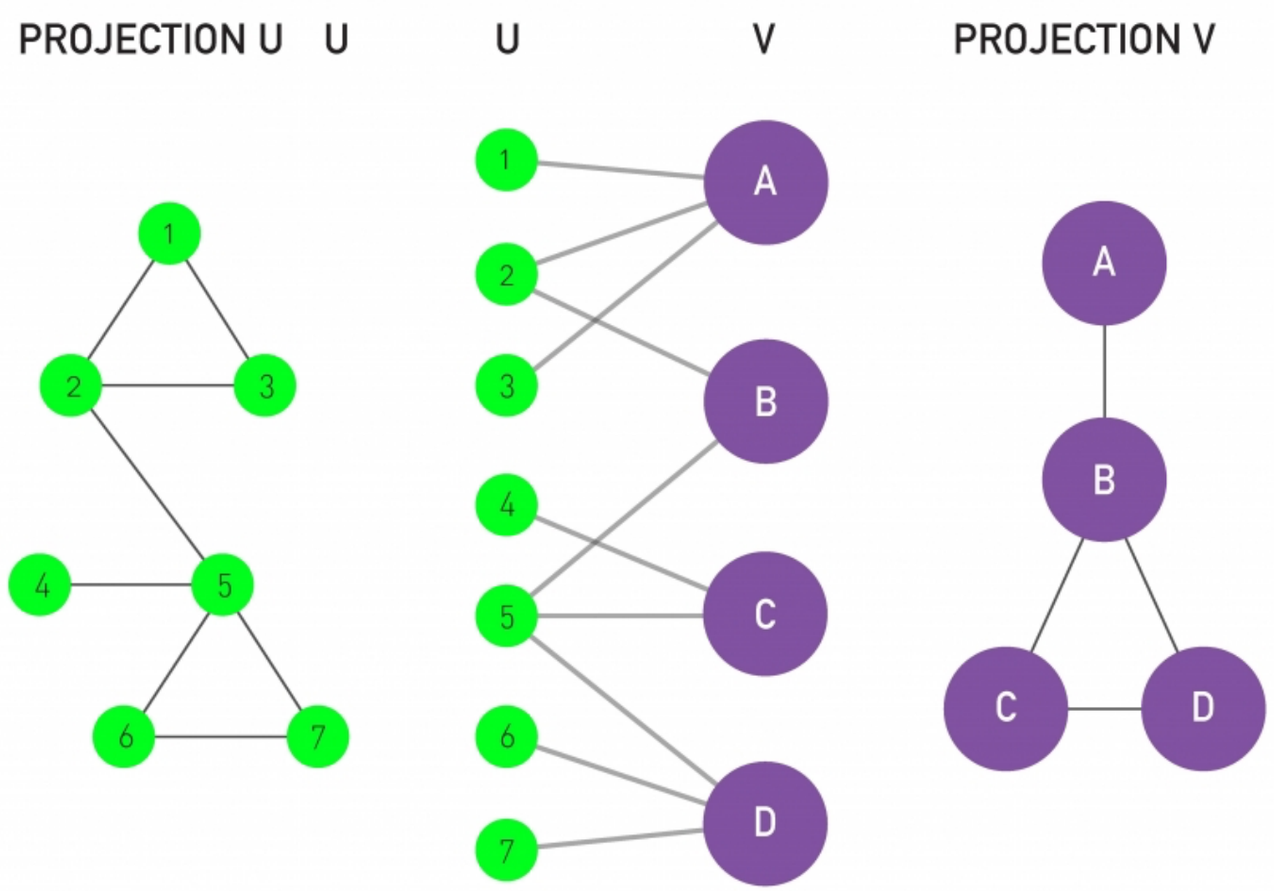
Mobile phone calls, email network

From [1]

Actors in the cast of the same movie

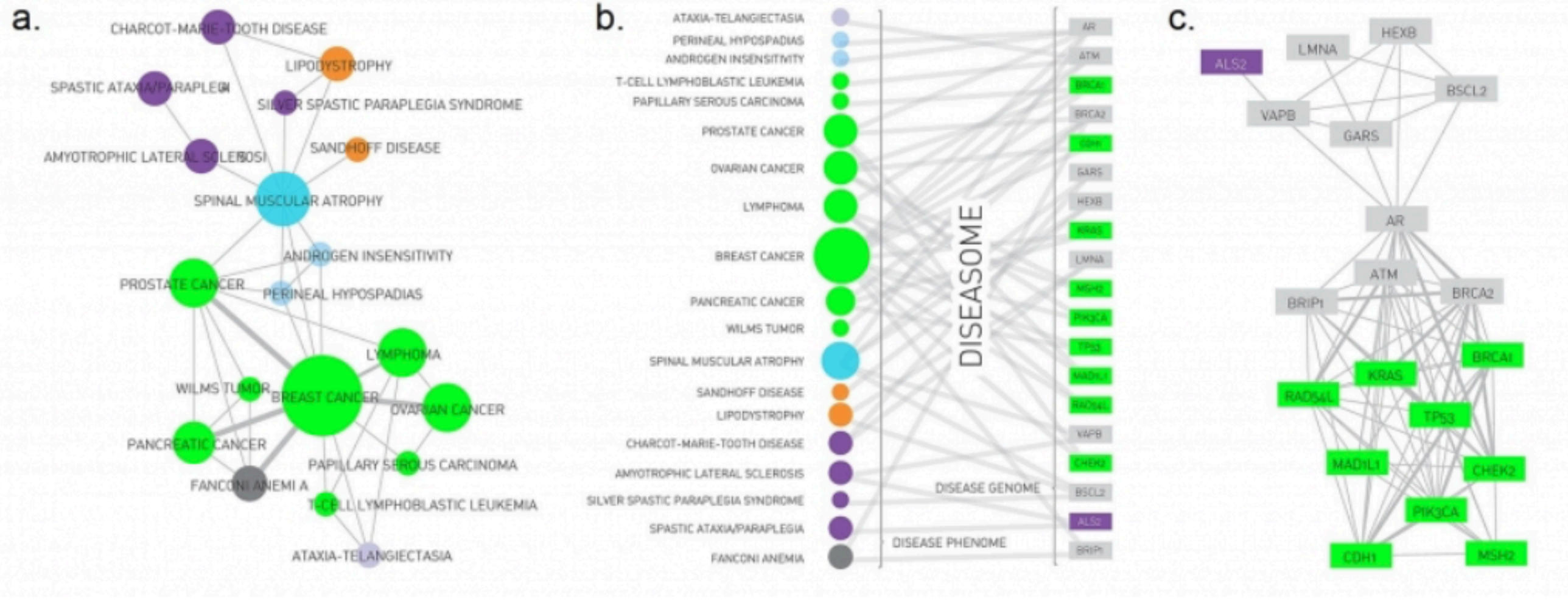
Special case: bipartite graphs

- A **bipartite graph** (or bigraph) is a graph whose nodes can be divided into two disjoint sets *U* and *V* such that every link connects a node in *U* to one in *V*; that is, *U* and *V* are independent sets



From [1]

Bipartite network: Human disease network



U projection: Disease network | **Bipartite network [diseases (U) and genes (V)]** | **V projection: Gene network**

From [1]

Overview

- Networks and graphs
- Definitions
- **Network density**
- Pathology
- Connectedness

Network density

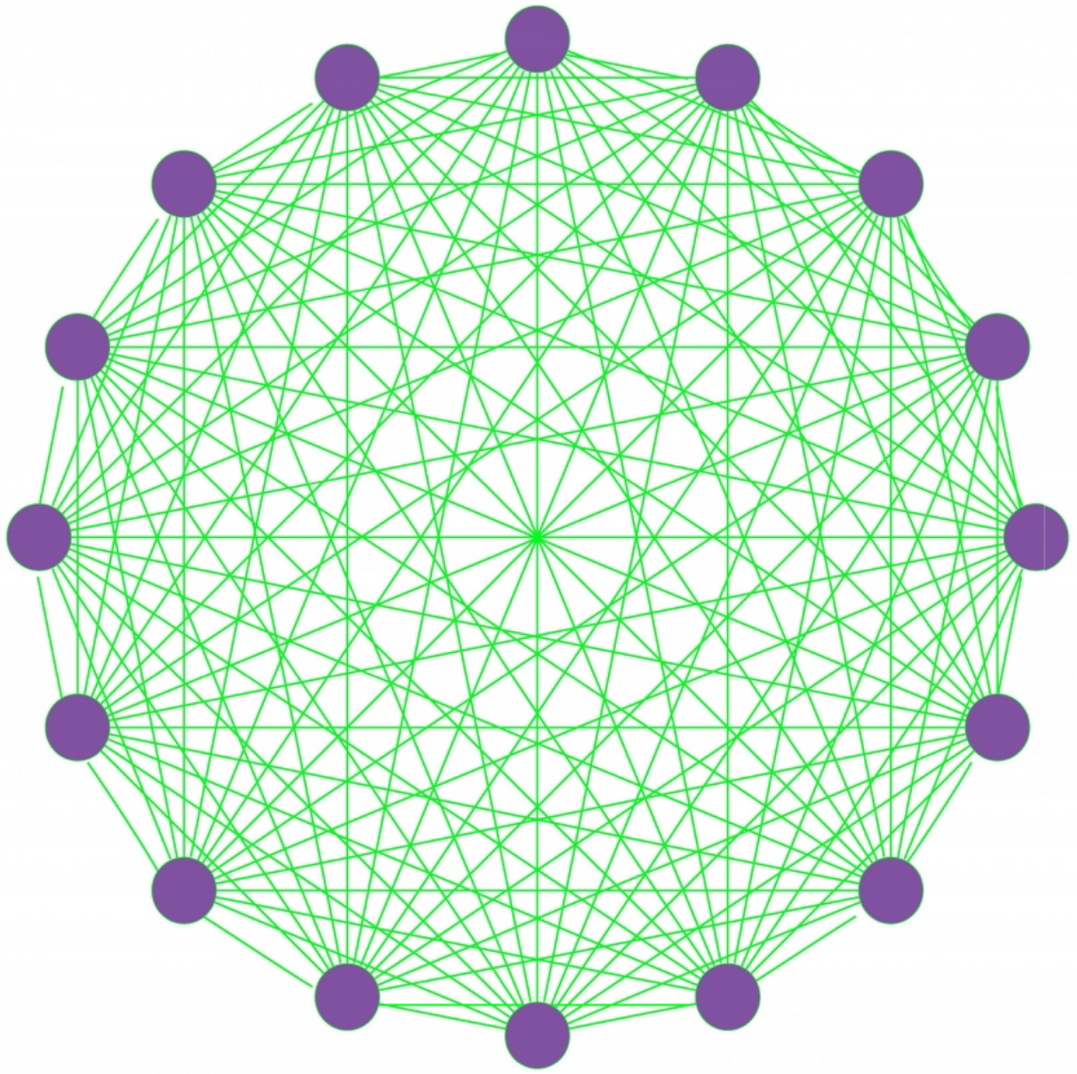
- Maximum number of links in a network:

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

  - a graph with $M = M_{max}$ is a *complete graph*, and $M = \mathcal{O}(N^2)$

  - a complete graph has an average degree of

$$\langle D \rangle = N - 1$$



From [1]

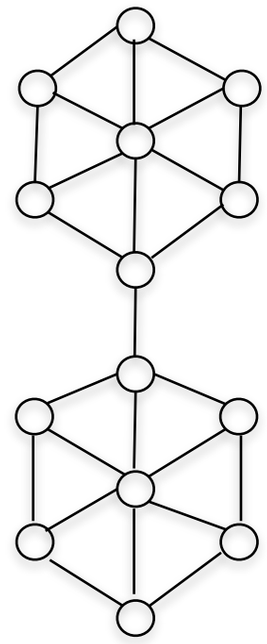
- Fortunately, most real networks are sparse:

$\langle D \rangle \ll N - 1$ and $M \ll M_{max}$, typically $M = \mathcal{O}(N)$

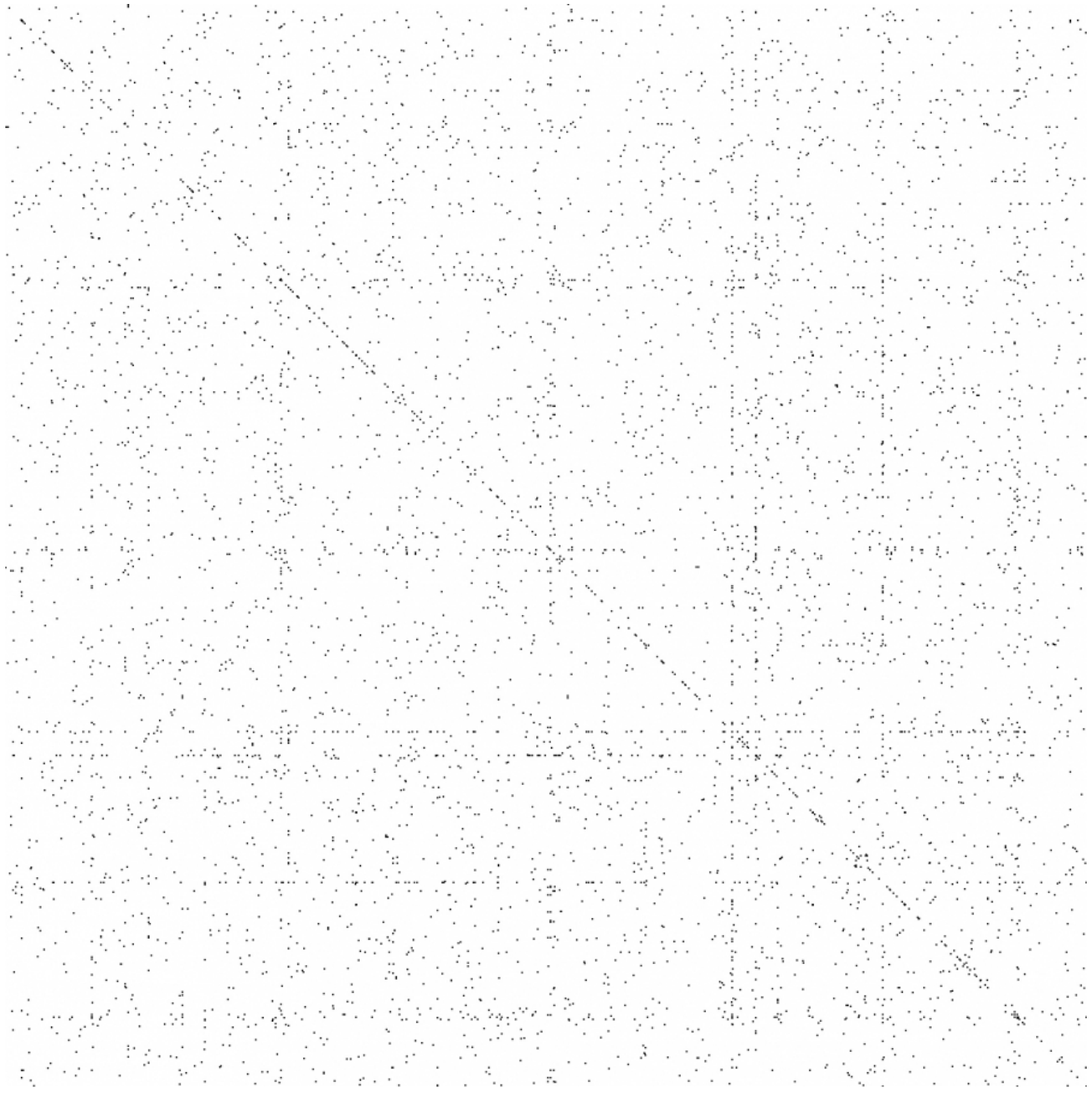
Real networks are sparse

- Sparse networks are highly desirable for memory and computational efficiency
- Good news: most natural/real-world networks (Facebook, Brain, Communication) are sparse
- Sparsity may relate to non-trivial structure

	N	M	M_{max}	$\langle D \rangle$
WWW (ND Sample)	325729	$1.4\ 10^6$	10^{12}	4.51
Protein (*S. Cerevisiae*)	1870	4470	10^7	2.39
Coauthorship (Math)	70975	$2\ 10^5$	$3\ 10^{10}$	3.9
Movie Actors	212250	$6\ 10^6$	$1.8\ 10^{13}$	28.78
Internet (2016)	$4.73\ 10^9$	10^{11}	10^{18}	



Adjacency matrix of real network



From [1]

protein-protein interaction network of yeast

Overview

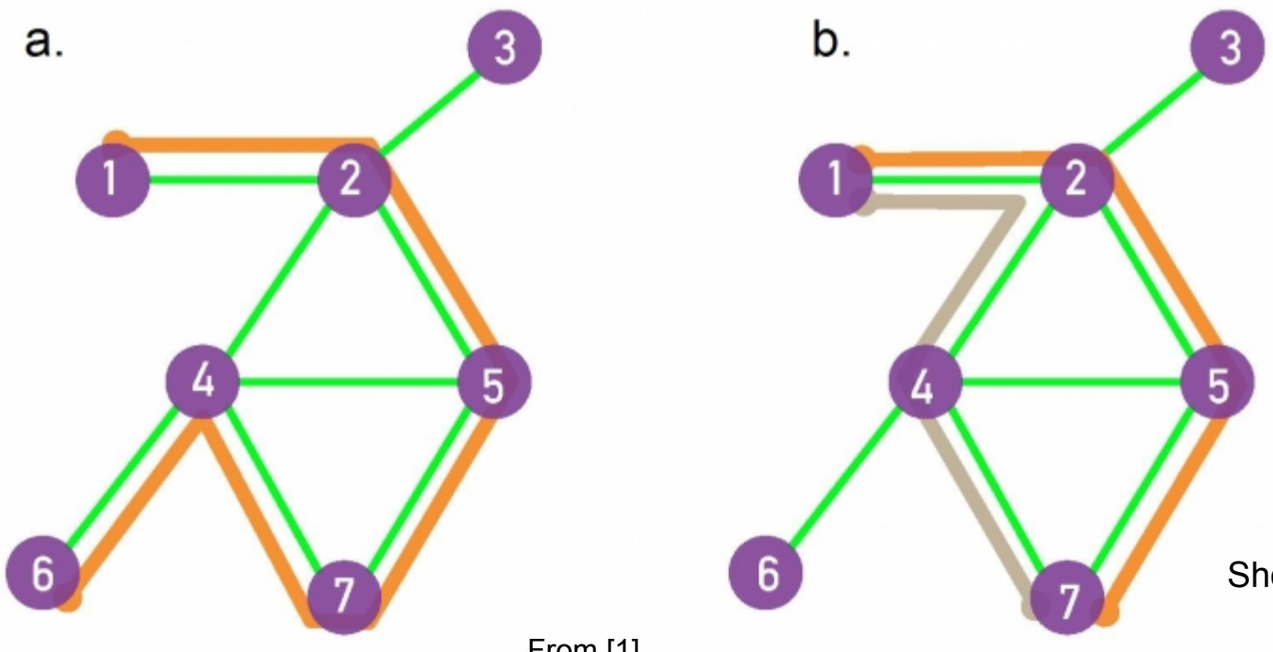
- Networks and graphs
- Definitions
- Network density
- **Pathology**
- Connectedness

Paths as a proxy for distances

- Physical distance is challenging: often measured by path length
- A path is a sequence of nodes in which each node is adjacent to the next one
- P_{i_0,i_n} of length n between nodes i_0 and i_n is an ordered collection of n links and $n+1$ nodes

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\} \qquad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \ldots, (i_{n-1}, i_n)\}$$

A path between 1 and 6



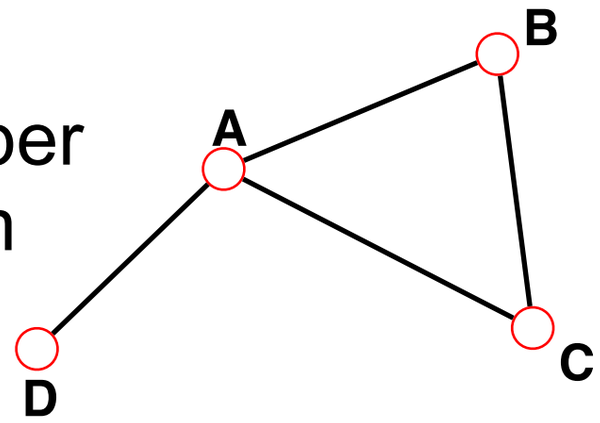
From [1]

Shortest paths between 1 and 7

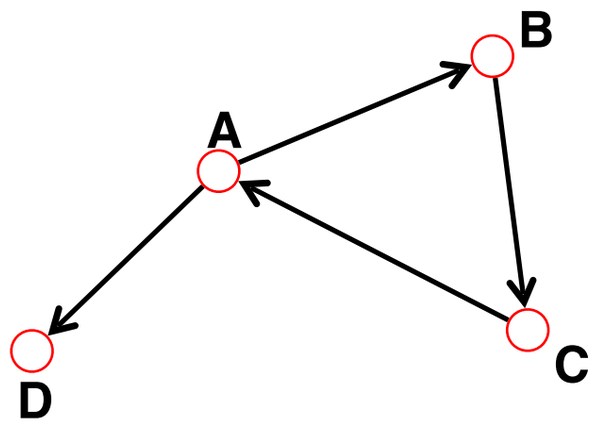
$$d_{17} = 3$$

Distance in a graph

- The **distance** d_{ij} (shortest path, geodesic path) between two nodes *i* and *j* is defined as the number of edges along the shortest path connecting them
  - If the two nodes are disconnected, the distance is infinity

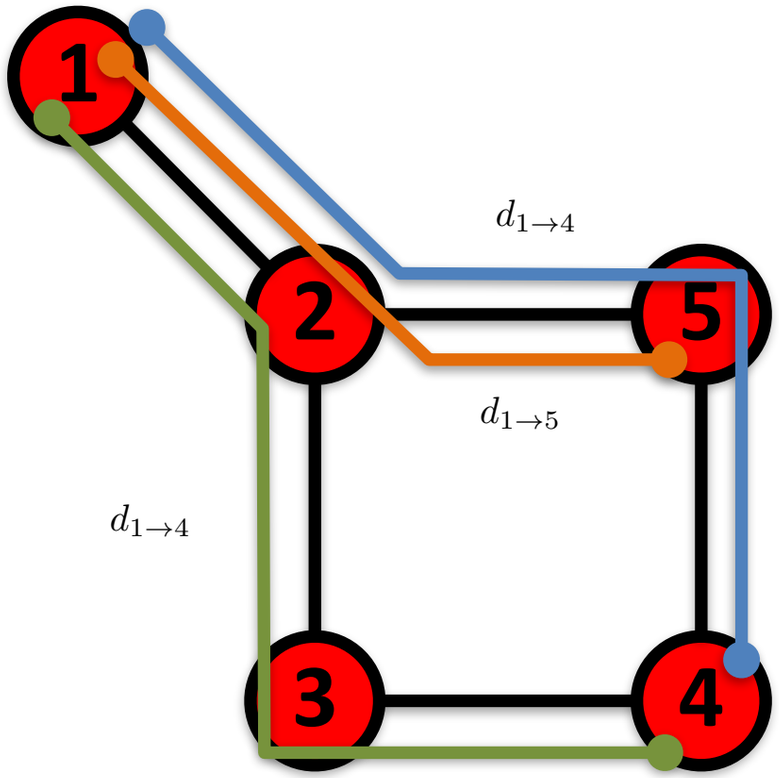


- In directed graphs each path needs to follow the direction of the arrows
  - Thus in a digraph the distance from node *i* to *j* is generally different from the distance from node *j* to *i*



- On weighted graphs, a weight distance can be defined in addition to the hop-distance

Shortest path



$d_{1\to4} = 3$

$d_{1\to5} = 2$

The path with the shortest length between two nodes (distance)

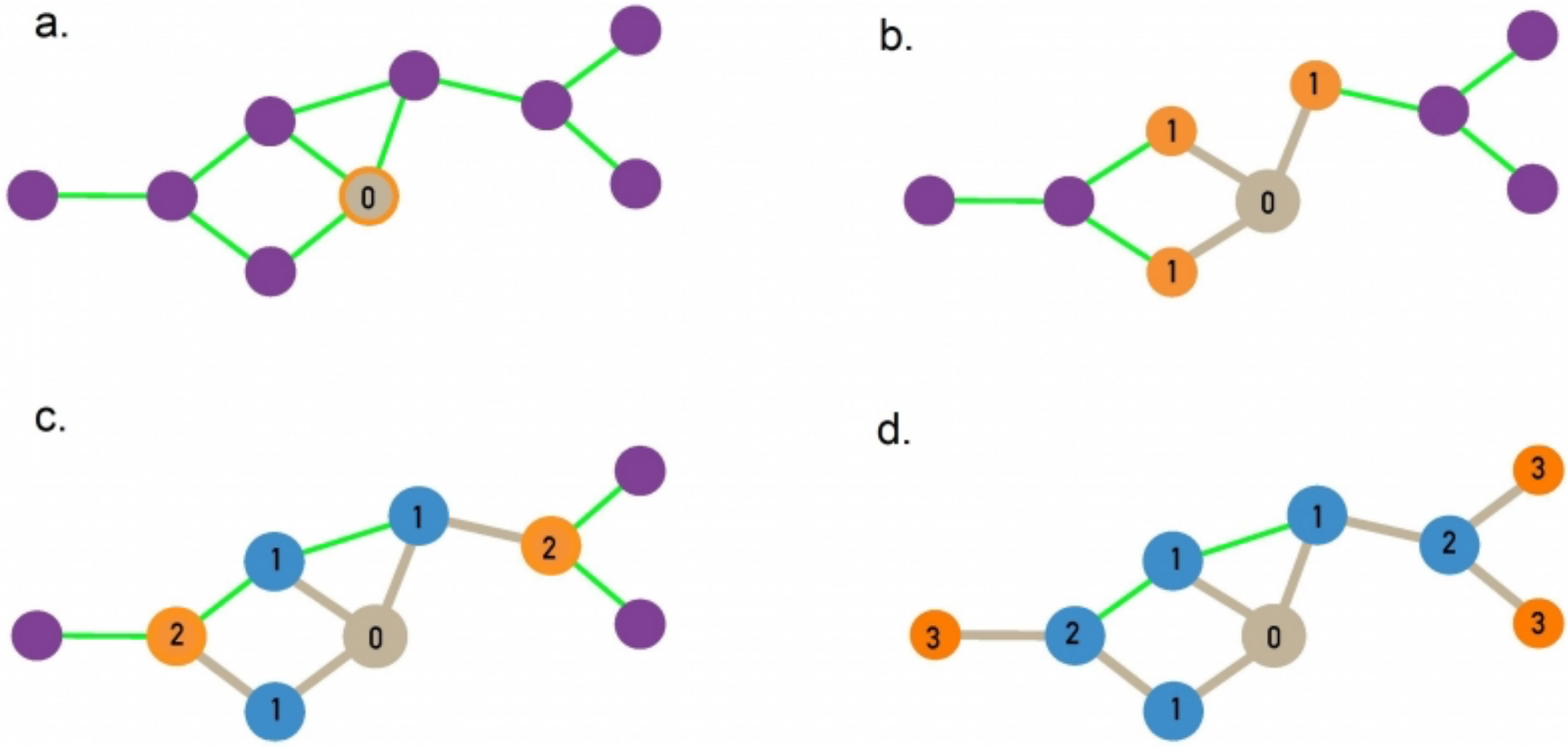
Finding shortest path: BFS

- Breadth-First Search (BFS): frequently used algorithm in network science for finding shortest path between *i* and *j*

Algorithm 1 BFS Algorithm: Shortest path between i and j

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

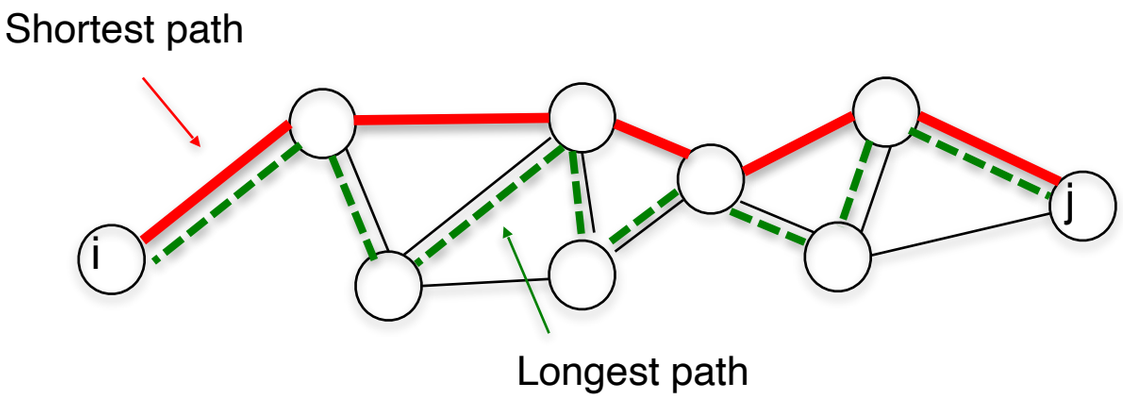
BFS algorithm illustration

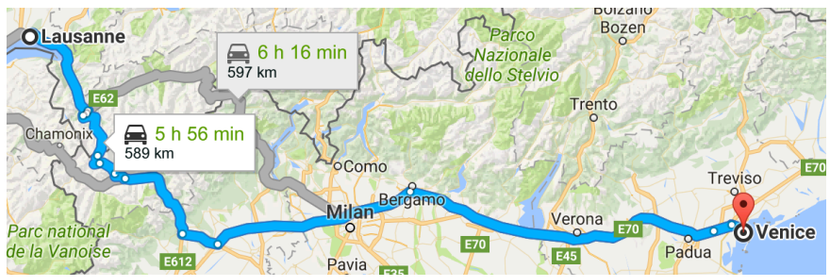


From [1]

Shortest path: a classic!

- Road navigation, e.g., Lausanne to Venice





- On weighted graphs, the shortest distance can be defined based on the sum of edge weights along the path, rather than the hop distance
  - Fast Algorithm: Dijkstra's algorithm (1956)

Number of paths between nodes

- We define with N_{ij} the number of paths between two nodes *i* and *j*

- Path of length $n = 1$: If there is a link between *i* and *j*, then $A_{ij} = 1$; Otherwise $A_{ij} = 0$
- Path of length $n = 2$: If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj} = 1$; Otherwise $A_{ik}A_{kj} = 0$
  - Number of paths of length two:

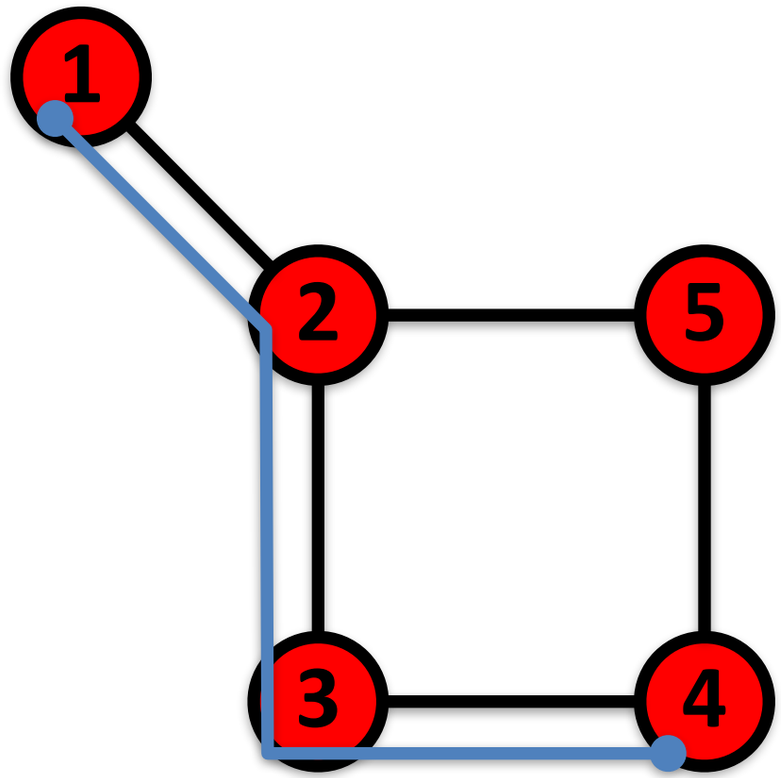
$$N_{ij}^{(2)} = \sum_{k=1}^{N} A_{ik}A_{kj} = \left[A^2\right]_{ij}$$

- Path of length *n*: in general, if there is a path of length *n* between *i* and *j*, then $A_{ik} \dots A_{lj} = 1$; Otherwise $A_{ik} \dots A_{lj} = 0$
  - Number of paths of length *n*:

$$N_{ij}^{(n)} = \left[A^n\right]_{ij}$$

Network diameter

- The network diameter d_{max} is the maximum distance between any pair of nodes in the graph (i.e., the longest shortest path)



$d_{1\to 4} = 3$

Average distance

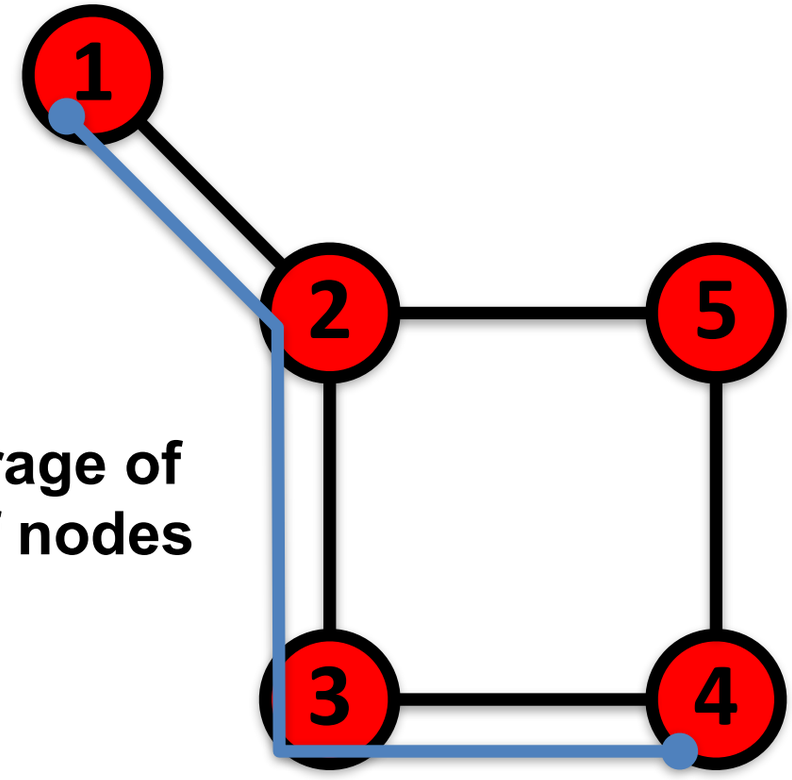
- The average path length (or average distance) for a connected graph is

$$\langle d \rangle \equiv \frac{1}{N(N-1)} \sum_{i,j \neq i} d_{ij} \quad \text{, or} \quad \langle d \rangle \equiv \frac{2}{N(N-1)} \sum_{i,j > i} d_{ij} \text{ if } \quad d_{ij} = d_{ji}$$

(undirected graphs)

  - with d_{ij} the distance between node *i* and *j*

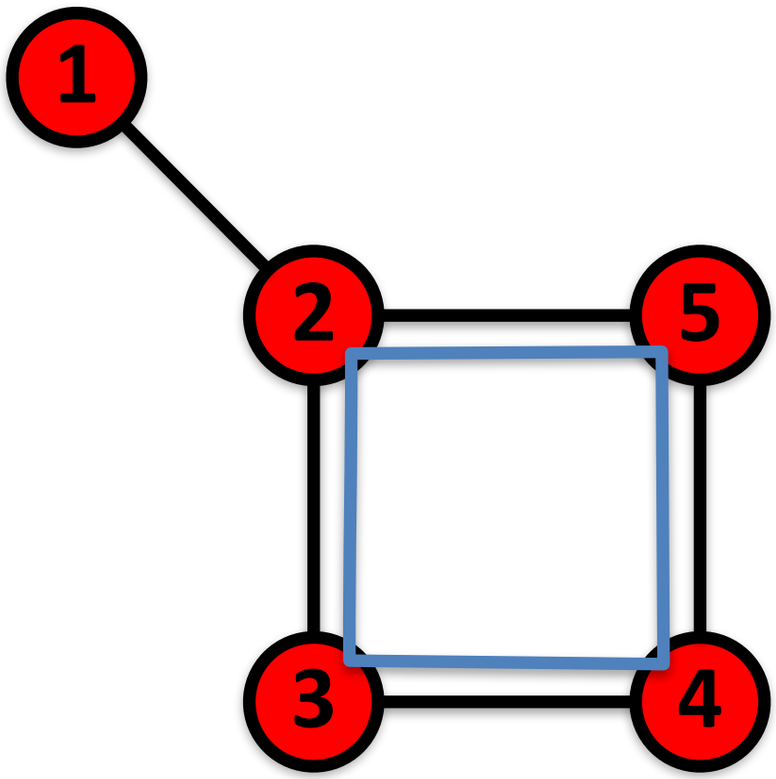
The average distance is the average of the shortest paths for all pairs of nodes



$$(d_{1\to2} + d_{1\to3} + d_{1\to4} + d_{1\to5} + d_{2\to3} + d_{2\to4} + d_{2\to5} + d_{3\to4} + d_{3\to5} + d_{4\to5})/10 = 1.6$$

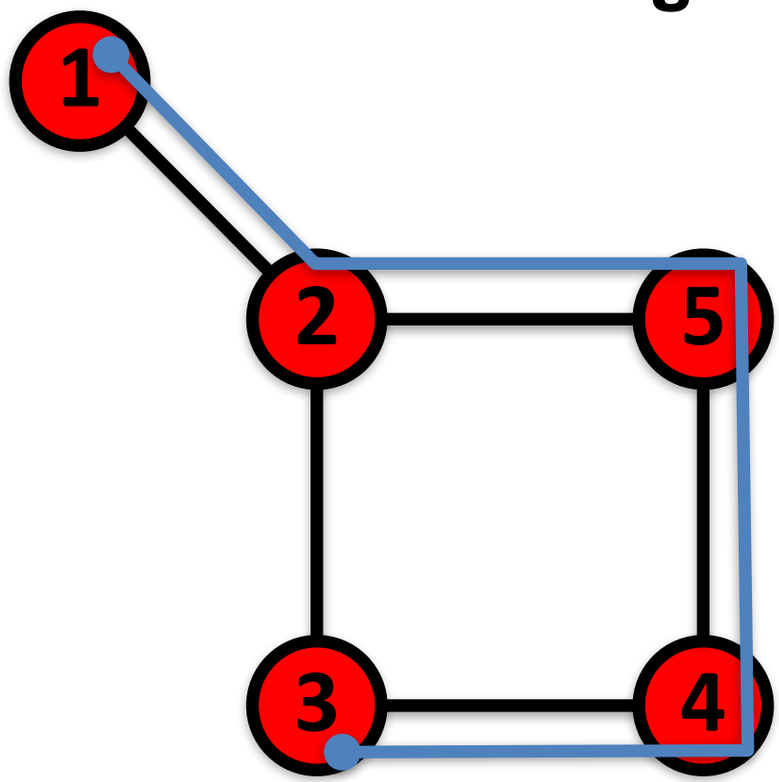
Other paths

Cycle



A path with the same start and end node

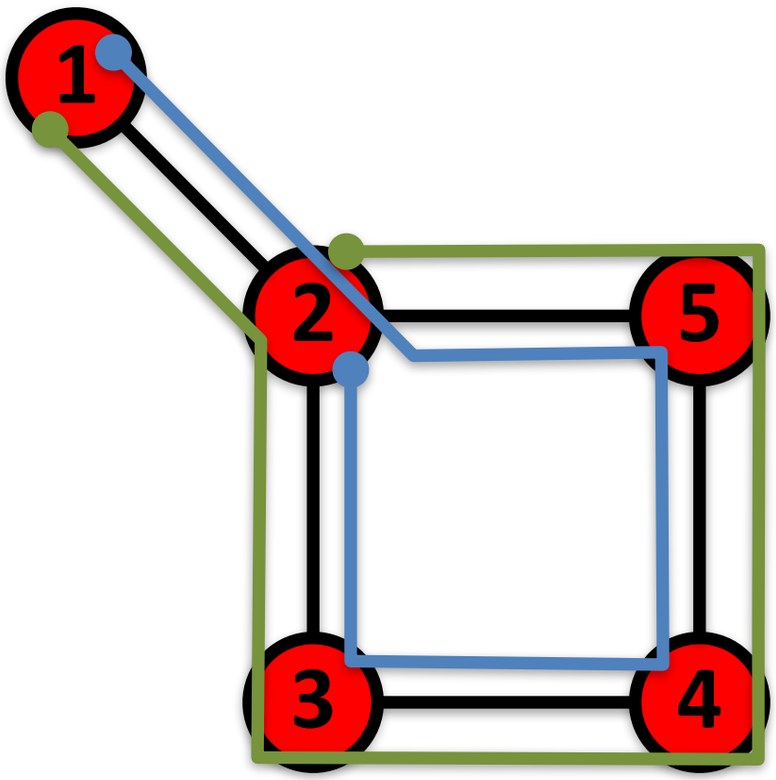
Self-avoiding Path



A path that does not intersect itself

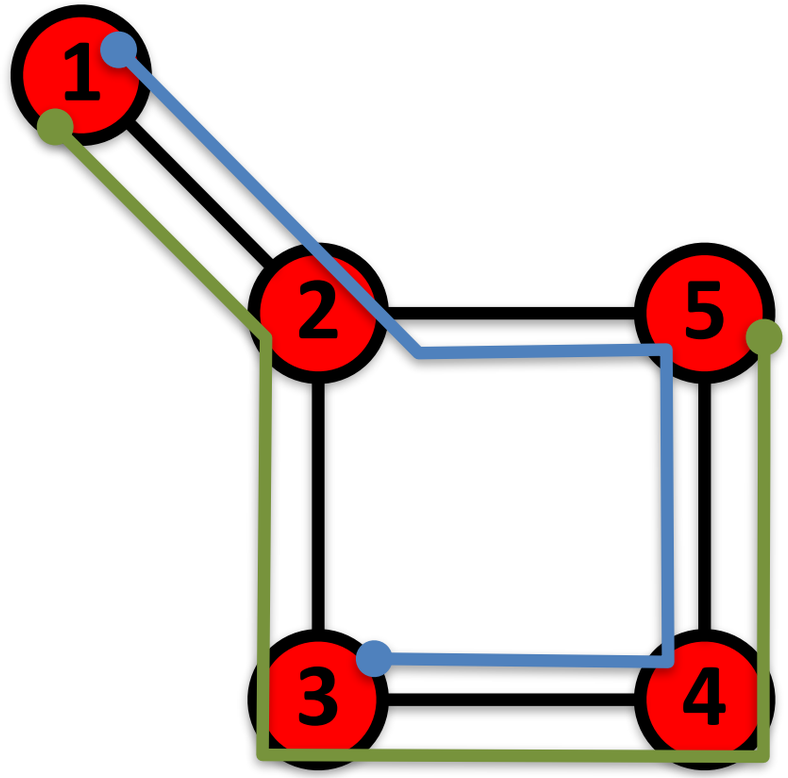
Yet more paths

Eulerian Path



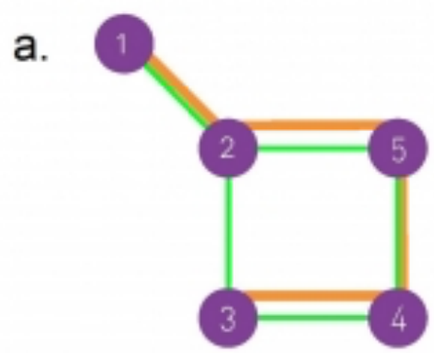
A path that traverses each link exactly once

Hamiltonian Path

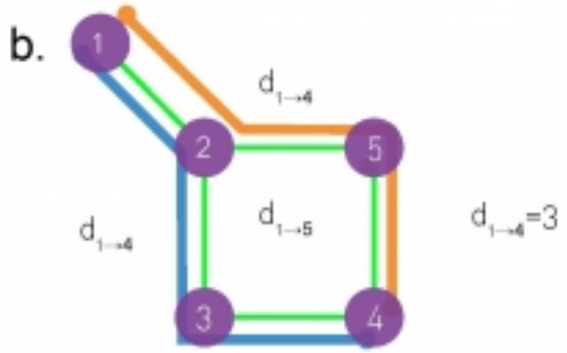


A path that visits each node exactly once

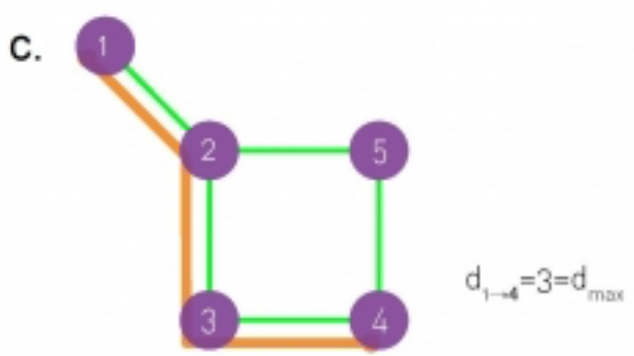
Pathology: summary



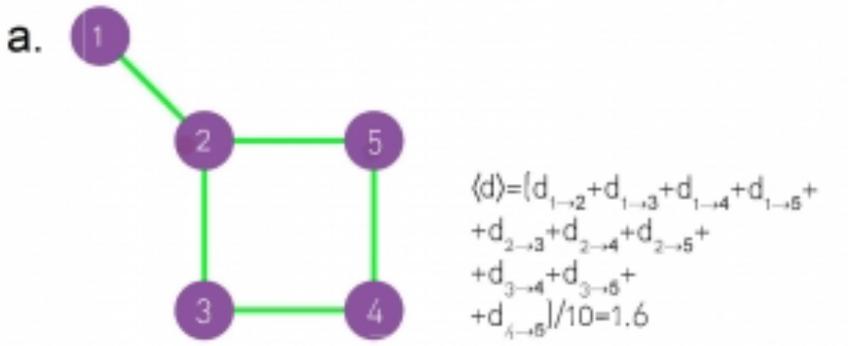
Path



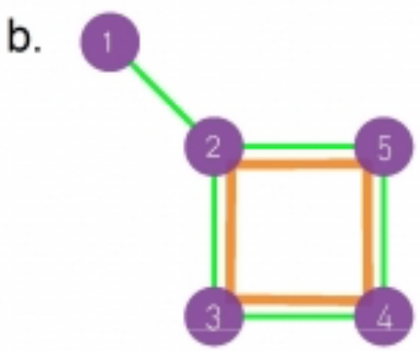
Shortest path



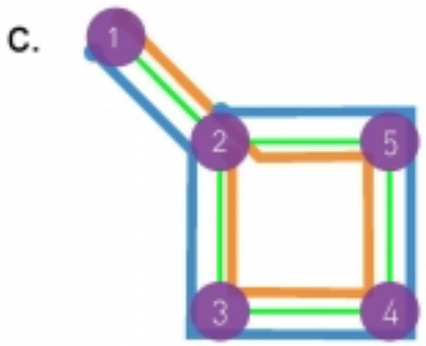
Diameter



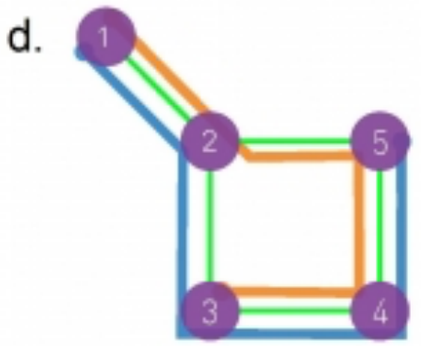
Average path length



Cycle



Euler path



Hamiltonian path

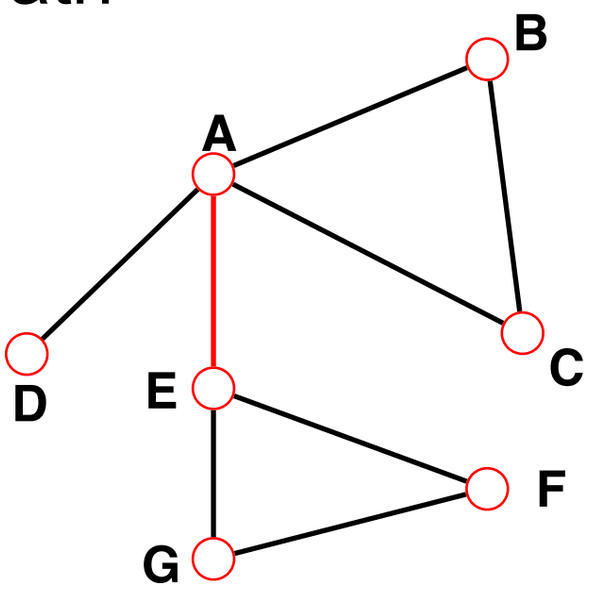
From [1]

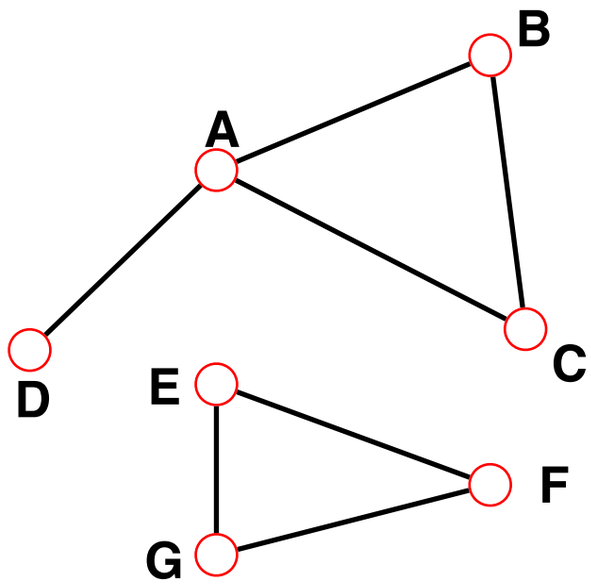
Overview

- Networks and graphs
- Definitions
- Network density
- Pathology
- **Connectedness**

Connected graphs

- Connectedness: a key property of networks
- A connected graph is a graph where any two vertices can be joined by a path





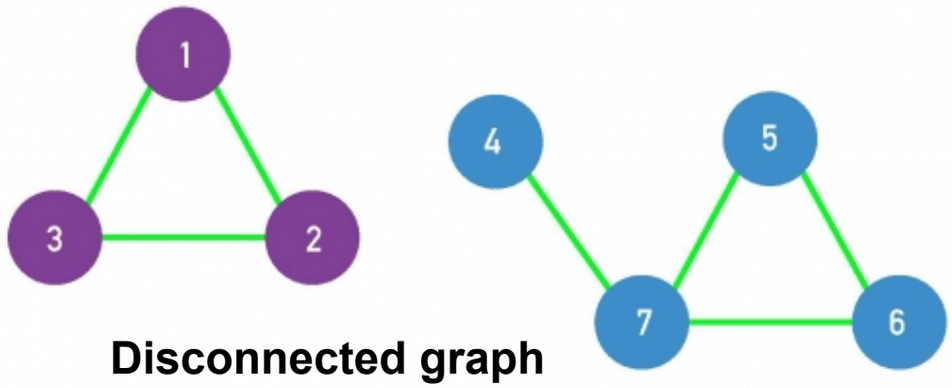
Largest Component: ***Giant Component***

The rest: ***Isolates***

- A disconnected graph is made up by two or more connected components/clusters
  - A bridge is a link that, if erased, leads to a disconnected graph

Connectivity of undirected graphs

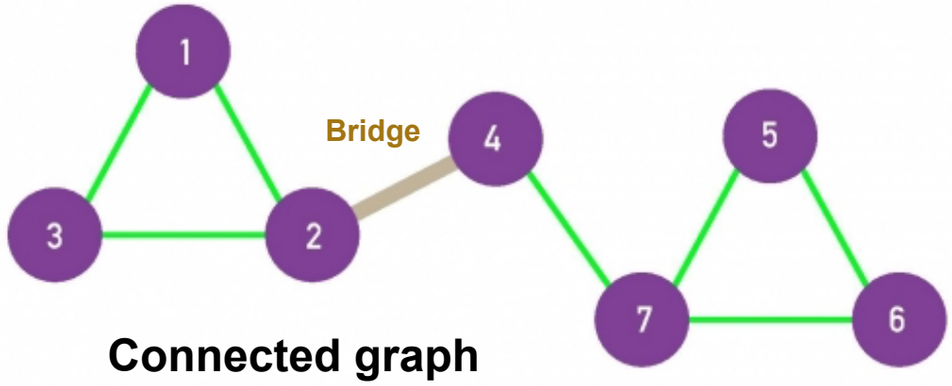
a.



Disconnected graph

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

b.



Connected graph

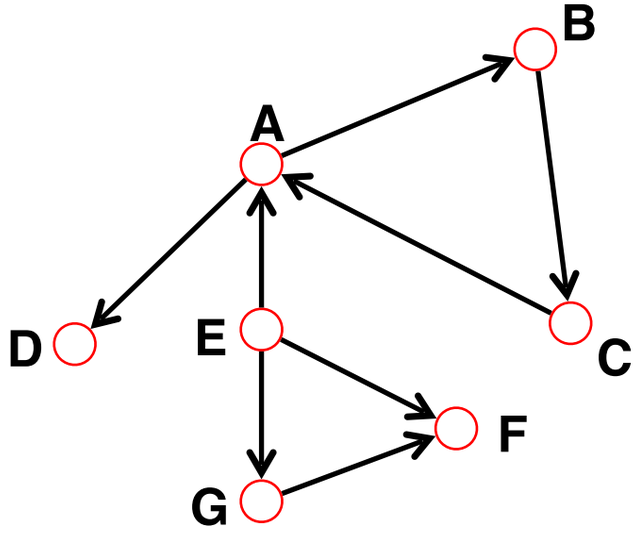
$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}$$

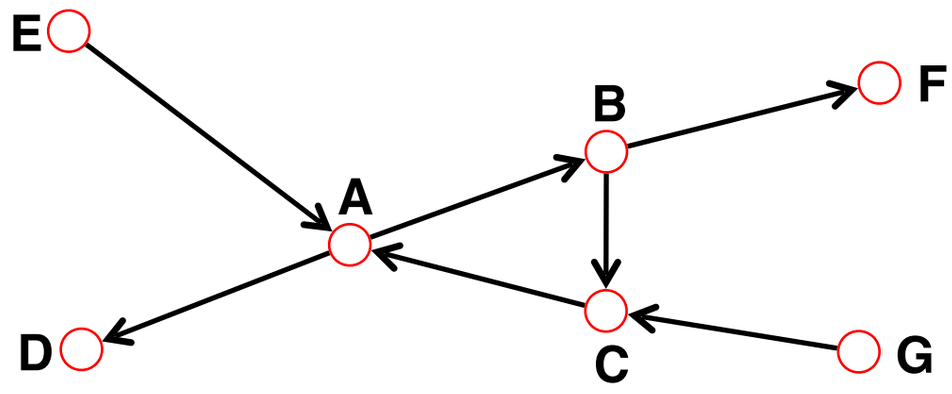
From [1]

- The adjacency matrix of a network with several components can be written in a block-diagonal form, so that nonzero elements are confined to squares, with all other elements being zero

Connectivity of directed graphs

- A **strongly connected** directed graph has a path from each node to every other node and vice versa (e.g., AB path and BA path)
- A **weakly connected** directed graph is a graph that is connected if we disregard the edge directions





Components identification

Algorithm 1 Finding connected components

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 labeled nodes equals N, then the network is connected. If the number of labeled 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

- The clustering coefficient is an indication about the fraction of the node's neighbors that are connected:

$$C_i = \frac{2M_i}{D_i(D_i-1)} \qquad \text{for} \qquad D_i \notin [0,1]$$

where M_i represents the number of links between the D_i neighbors of node *i*

- It represents the probability that two neighbors of a node, are linked to each other,

$$C_i \in [0,1]$$

- The average clustering coefficient is given by

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

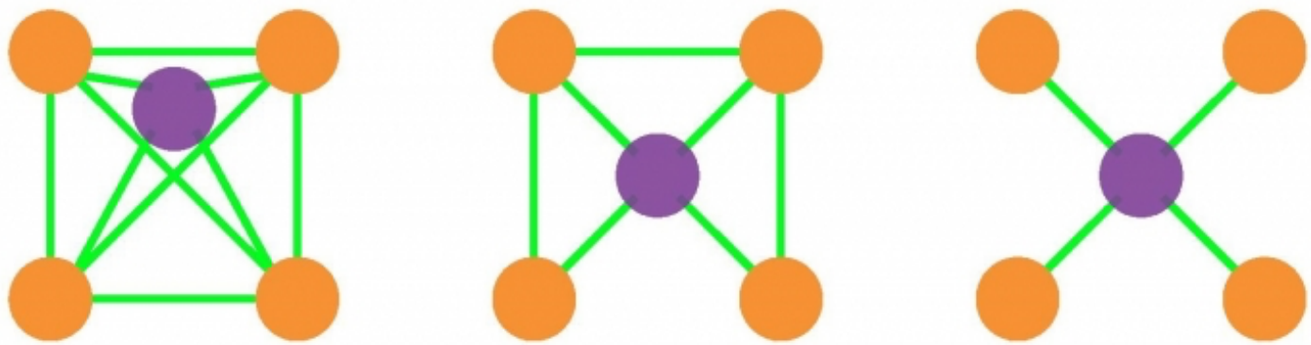
Global clustering coefficient

$$C_\Delta = \frac{3 \times \mathrm{NumberOfTriangles}}{\mathrm{NumberOfConnectedTriplets}}$$

- It measures the total number of closed triangles in a network
  - each link between two neighbours of node *i* closes a triangle
- It is normalised by the number of triplets
  - a triplet is a series of 3 connected nodes (not necessarily forming a triangle)
  - a triangle contains 3 triplets
- It is not identical to the average clustering coefficient
  - they generally have comparable values

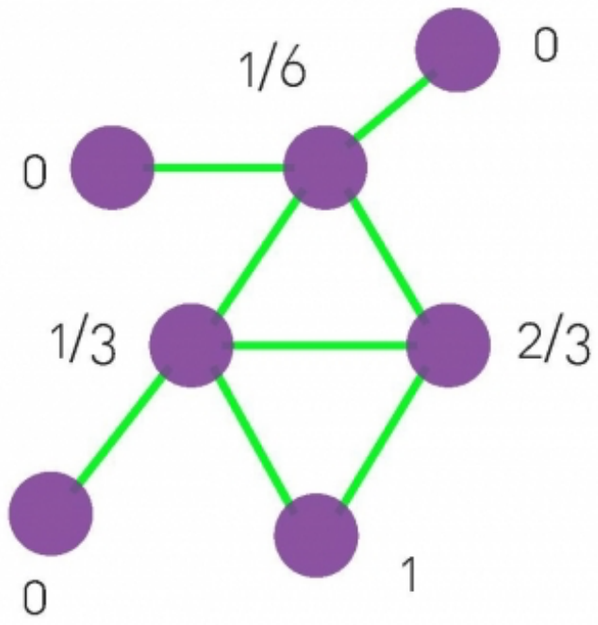
Clustering coefficient: examples

a.



$C_i=1$ $C_i=1/2$ $C_i=0$

b.

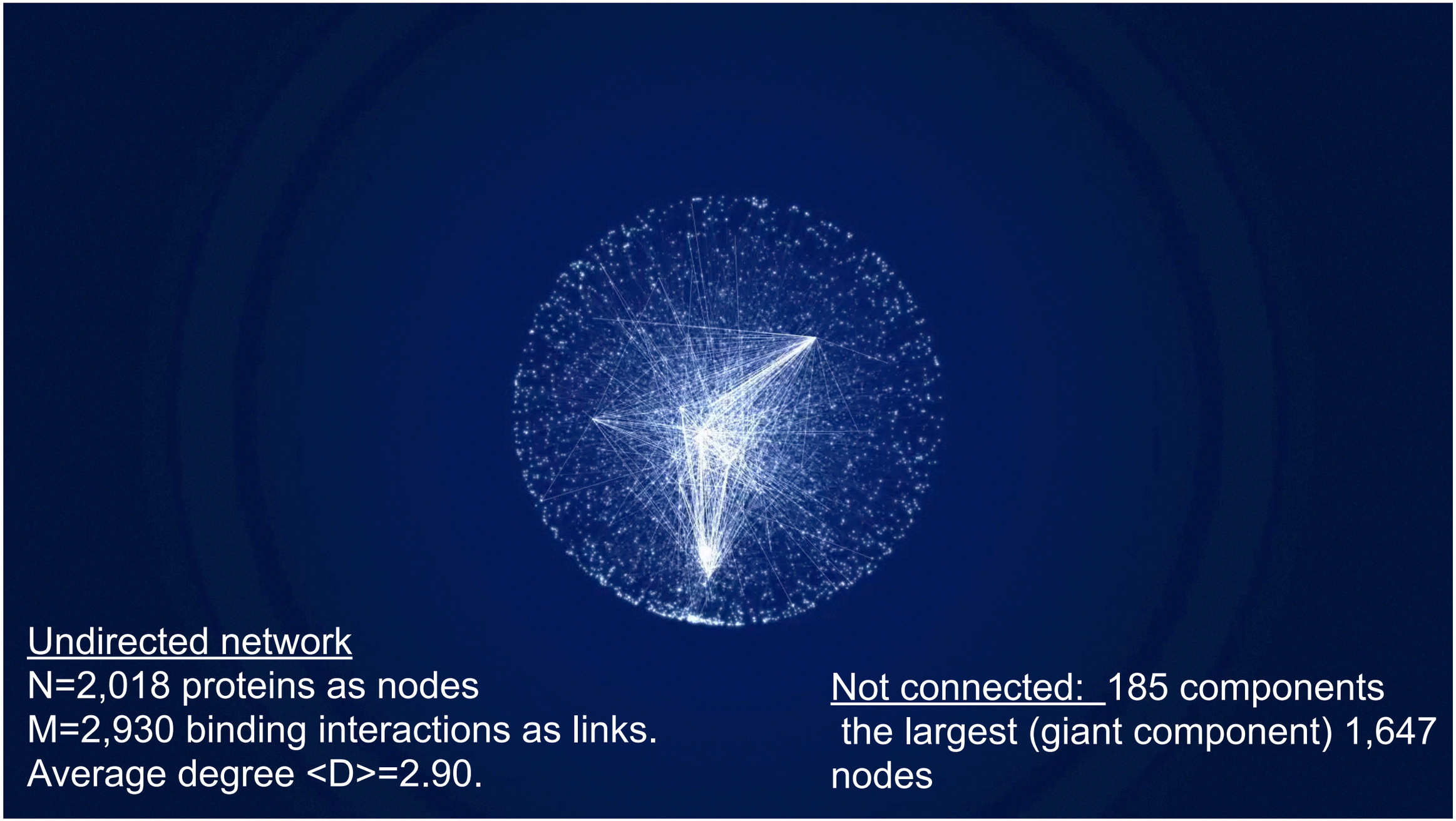


$$\langle C\rangle = \frac{13}{42} \approx 0.310$$

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

From [1]

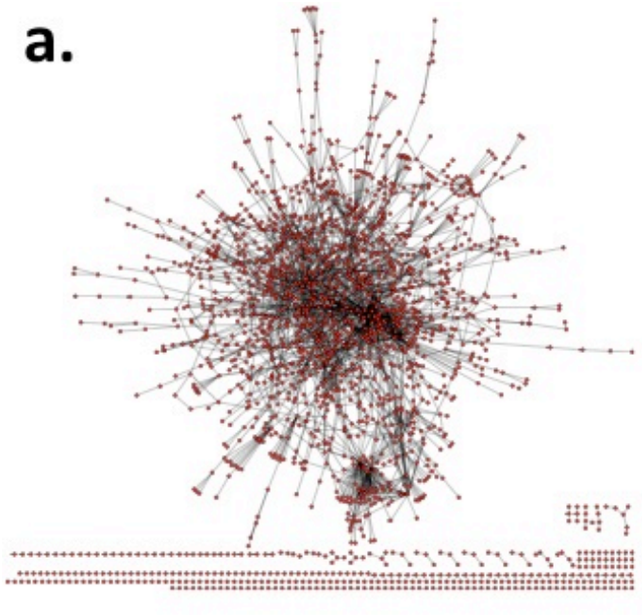
A case study: protein-to-protein interaction network

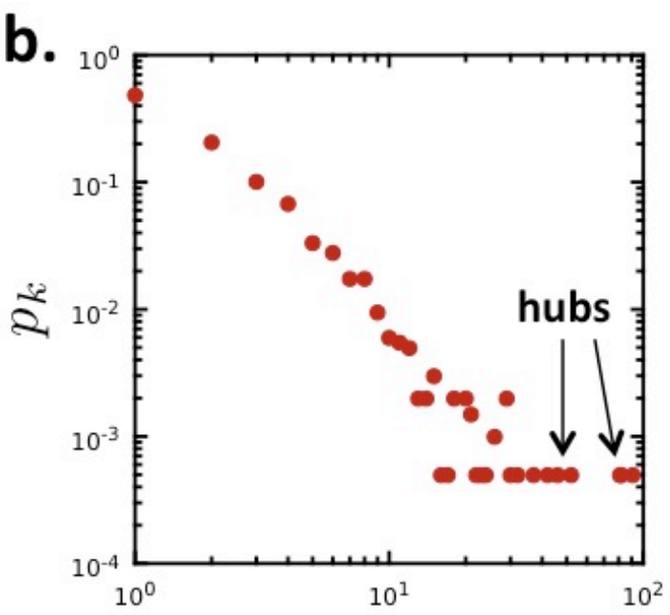


Undirected network
N=2,018 proteins as nodes
M=2,930 binding interactions as links.
Average degree <D>=2.90.

Not connected: 185 components
the largest (giant component) 1,647 nodes

Analysing central quantities



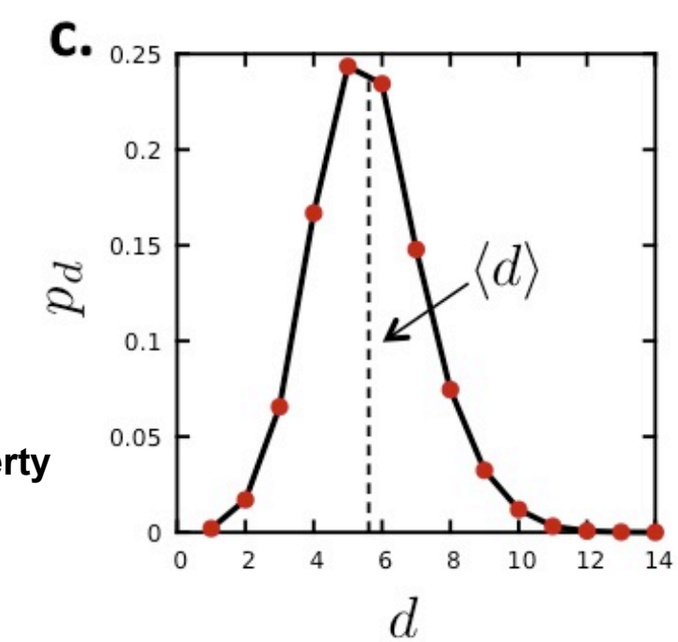


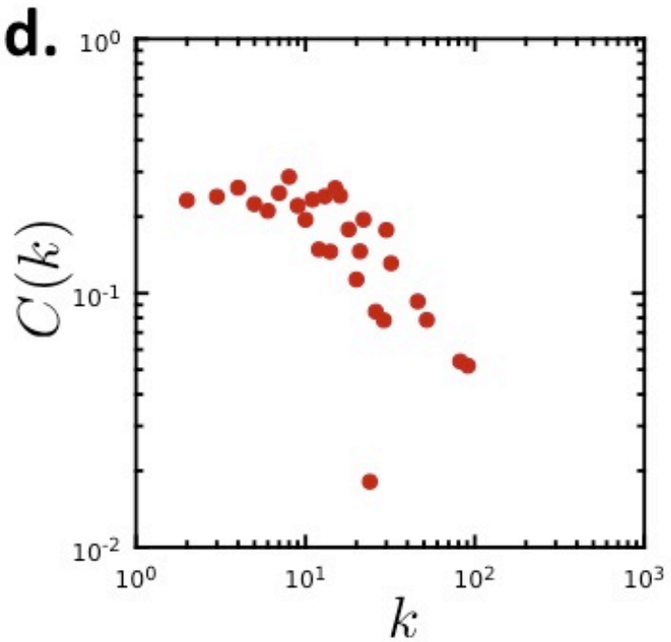
$$p_k = \frac{N_k}{N}$$

Scale free property

$$d_{max} = 14$$
$$\langle d \rangle = 5.61$$

Small world property





$$\langle C \rangle = 0.12$$

Network hierarchy

Summary

- A network is a combination of system's components often called nodes/vertices, and the direct interaction between them, called links/edges

- Networks can be represented as graphs

- Graphs can be natural or constructed

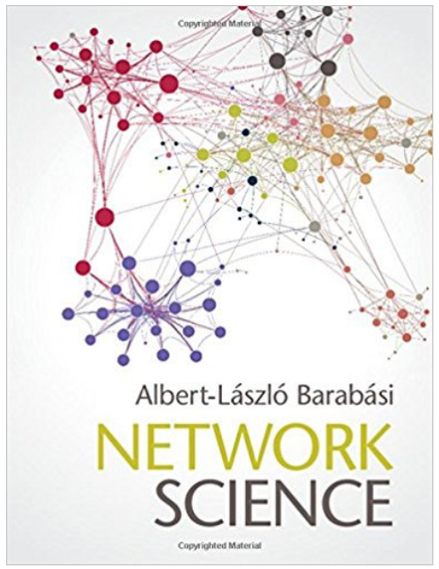
- Each graph has some characteristics:
  - Directed, undirected, weighted, unweights, bipartite, connected, disconnected etc

- Paths are proxies for distances on the graph

References

[1] Network Science, by Albert-László Barabási

- http://networksciencebook.com/chapter/2#bridges



Some content of the slides is inspired from Prof. Xavier Bresson's class 'A Network Tour of Data Science (2016), and from Prof. Barabási's class on Network Science (www.BarabasiLab.com)