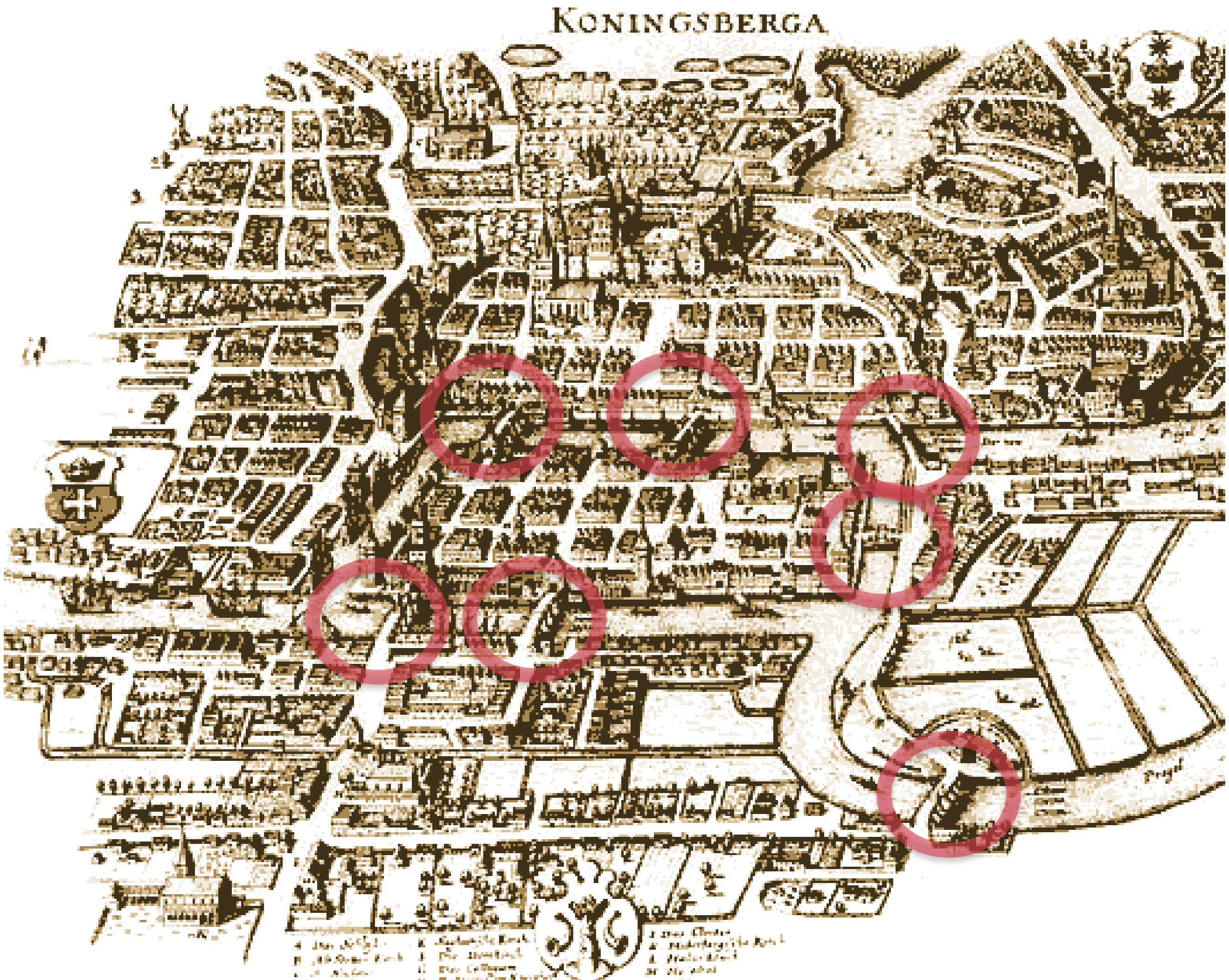
Graph Theory

Alberto Paccanaro
EMAp – FGV

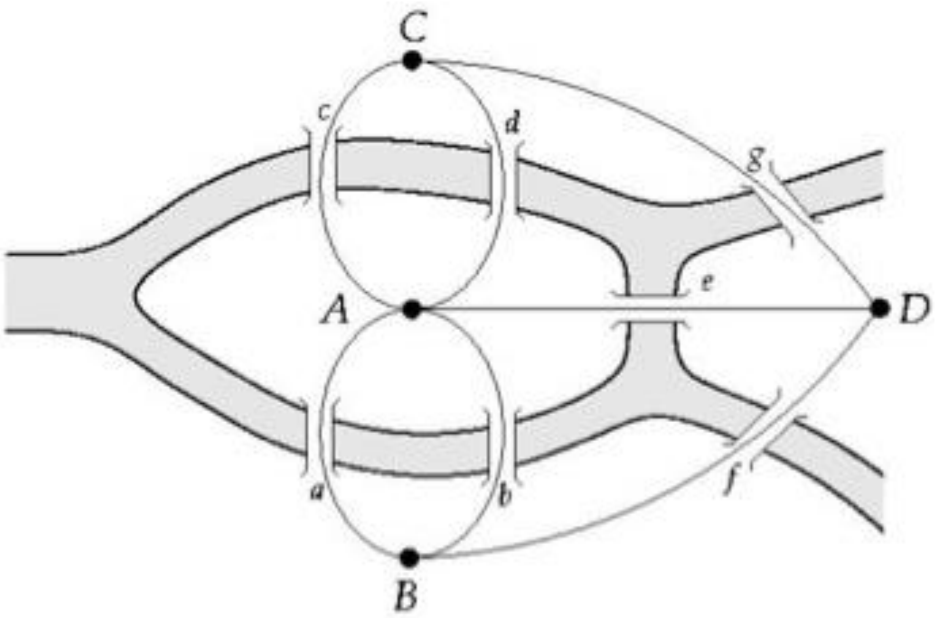
www.paccanarolab.org

Some material and images are from (or adapted from):
A. Barabási, and M. Pósfai. Network science, Cambridge University Press, 2016

The Bridges of Konigsberg



1735 Euler
Can one walk across the seven bridges and never cross the same bridge twice?



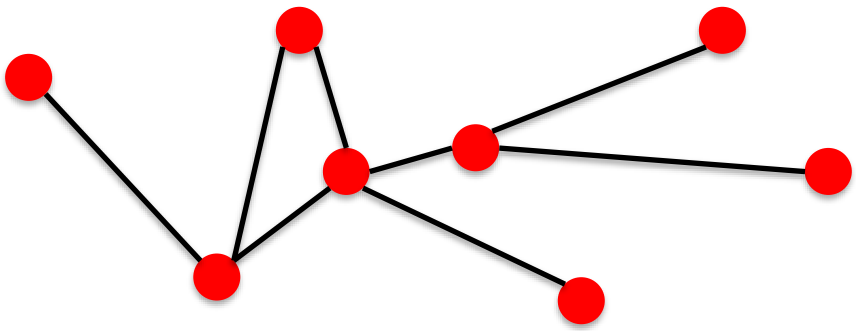
Euler's theorem (1735):

- 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.

1. Some problems become more treatable if they are represented as a graph (abstraction).

2. The existence of the path is a property of the graph.

Basic definitions



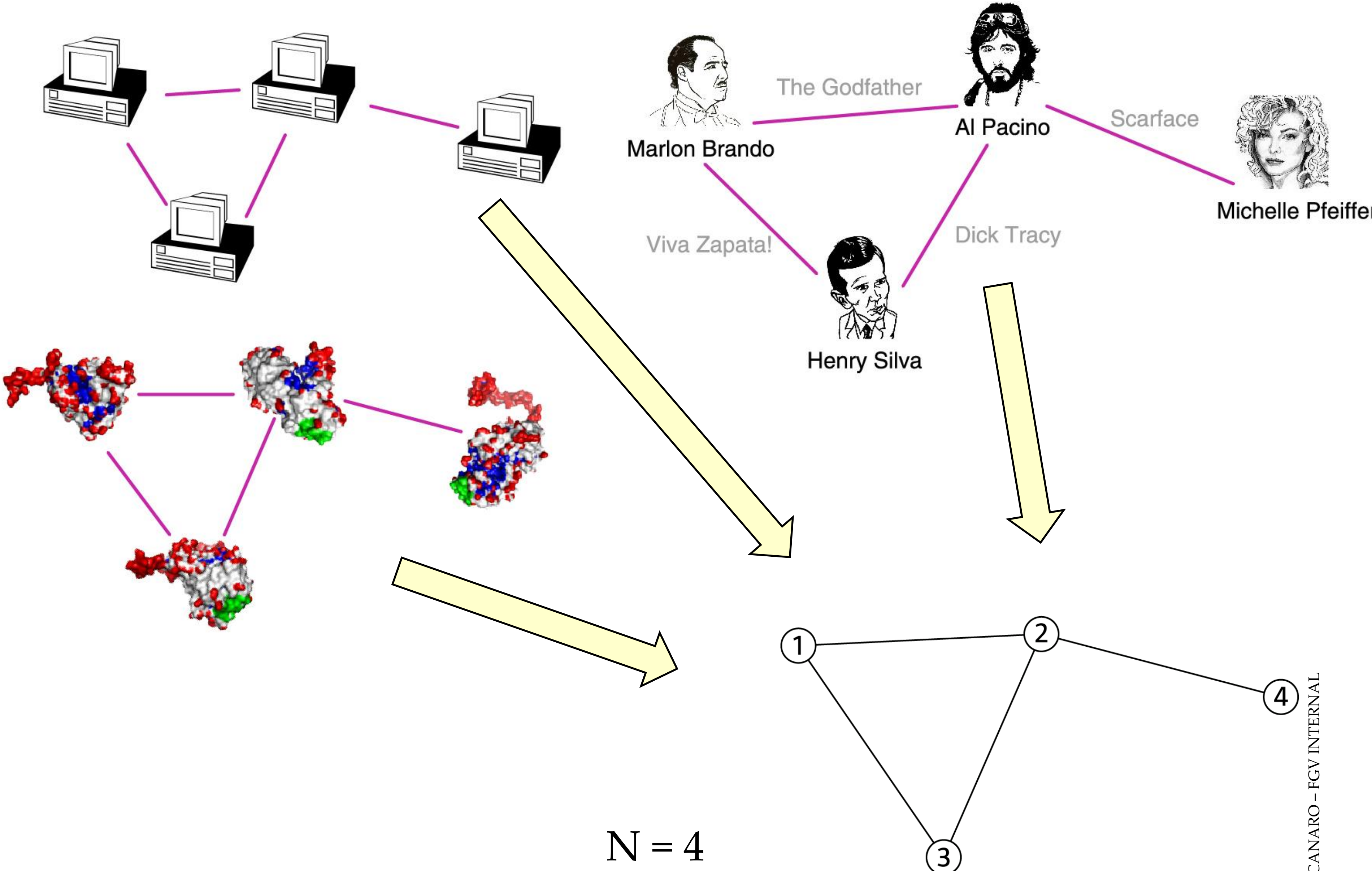
Nodes, vertices – ***N***
Links, edges – ***L***

(Network, node, link)

Network: refers to real systems (www, social network, metabolic network)

(Graph, vertex, edge)

Graph: mathematical representation of a network (web graph, social graph)


The Godfather
Marlon Brando
Al Pacino
Scarface
Michelle Pfeiffer
Viva Zapata!
Dick Tracy
Henry Silva
1
2
3
4

$N = 4$

$L = 4$

Remember…

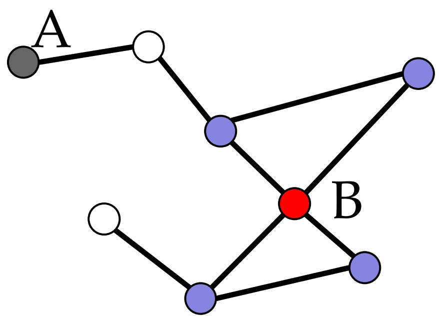
- The choice of the network representation determines our ability to use network theory successfully.

- In many cases, the representation is by no means unique.

- This choice will determine **the question** we can study.

Node Degree, k

Undirected

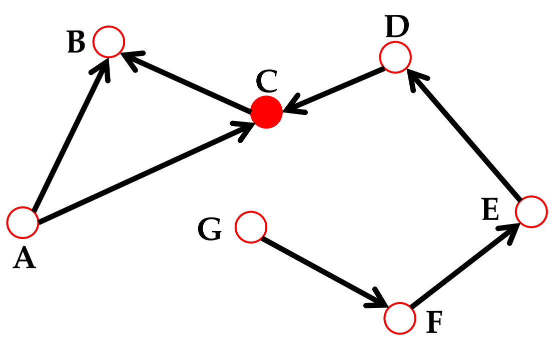


Node degree: the number of links connected to the node.

$$k_B = 4$$

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

Directed



Directed networks: **in-degree** and **out-degree**.

The (total) degree is the sum of in- and out-degree.

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

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

Source: a node with $k^{in} = 0$

Sink: a node with $k^{out} = 0$

Some stats…

BRIEF STATISTICS REVIEW

Four key quantities characterize a sample of N values $x_1, \ldots, x_N$:

Average (mean):

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

The n^{th} moment:

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

Standard deviation:

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

Distribution of x:

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

where p_x follows

$$\sum_i p_x = 1 \left(\int p_x \, dx = 1\right)$$

Average Degree

Undirected:

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

Directed:

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

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

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

N – the number of nodes in the graph

Examples used in the Barabasi book

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-authorship	Undirected	23,133	93,439	8.08
Actor Network	Actors	Co-acting	Undirected	702,388	29,397,908	83.71
Citation Network	Paper	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

Degree distribution

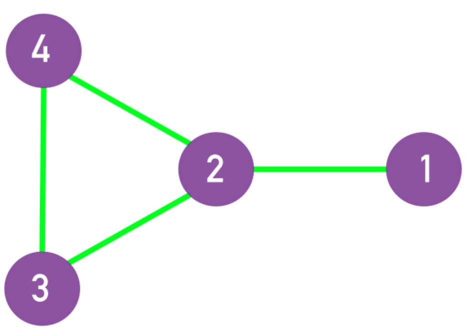
The degree distribution, p_k, provides the probability that a randomly selected node in the network has degree k.

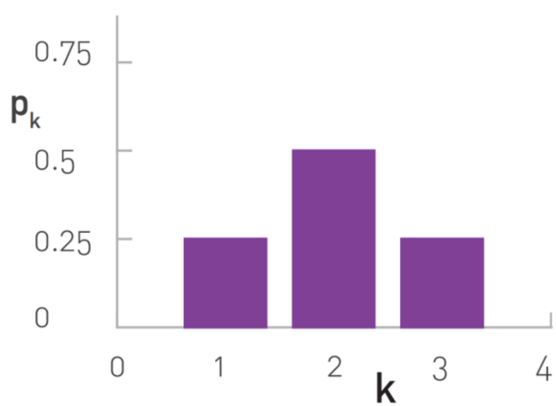
$$\boxed{p_k = \frac{N_k}{N}}$$

The form of p_k determines many phenomena

N_k number of nodes of degree k

It is a normalized histogram $\sum_{k=1}^{\infty} p_k = 1$

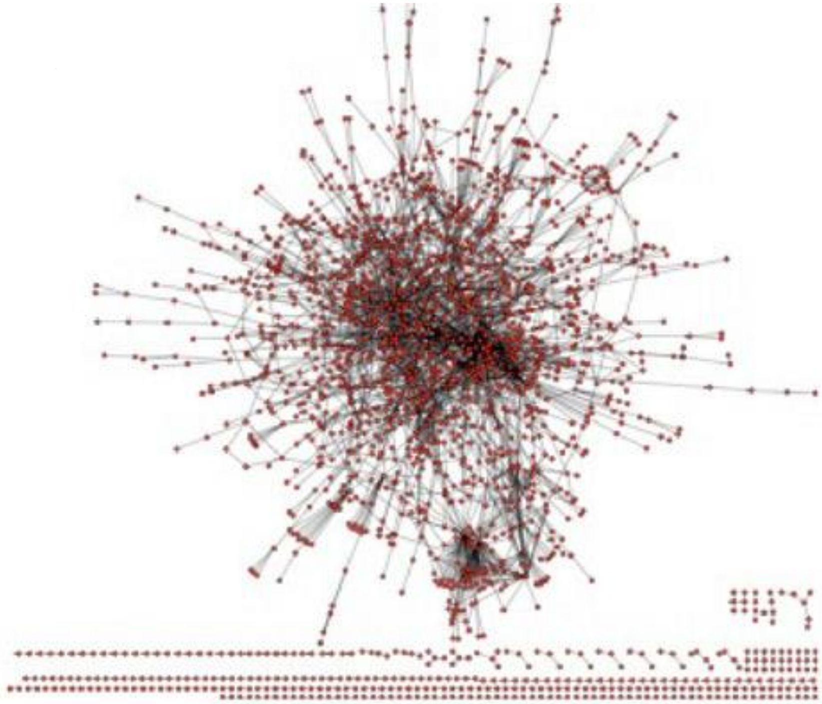


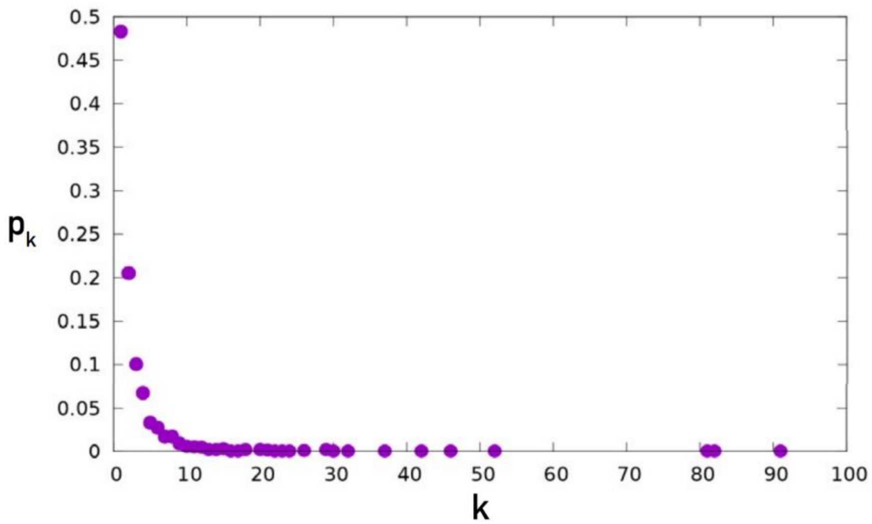


$$\langle k \rangle = \sum_{k=0}^{\infty} k p_k$$

$$N_k = N p_k$$

A real world example – protein protein interaction network





Adjacency matrix

We consider networks without self-loops

Undirected network: binary matrix N x N

- $A_{ij} = A_{ji} = 1$ if i is connected to j
- $A_{ij} = 0$ otherwise

$$\boxed{k_i = \sum_{j=1}^{N} A_{ji} = \sum_{i=1}^{N} A_{ji}}$$

$$\boxed{L = \frac{1}{2}\sum_{ij}^{N} A_{ij}}$$

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

Directed network: binary matrix N x N

- $A_{ij} = 1$ if there is a link from j to i
- $A_{ij} = 0$ otherwise

from

to $\boxed{A_{ij}}$

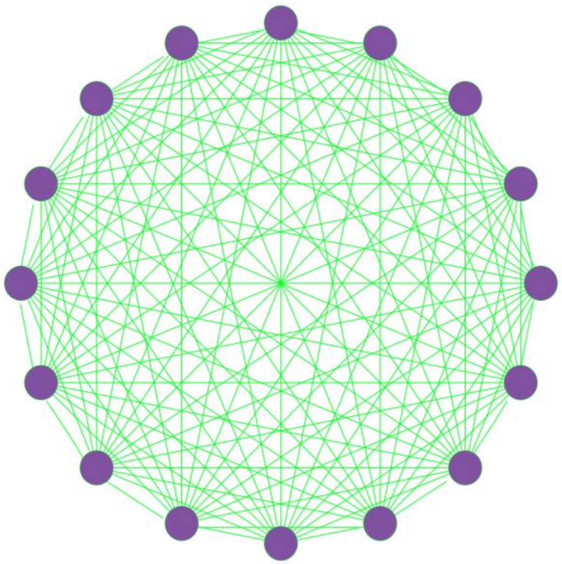
$$\boxed{k_i^{\text{in}} = \sum_{j=1}^{N} A_{ij}} \qquad \boxed{k_i^{\text{out}} = \sum_{j=1}^{N} A_{ji}} \qquad \boxed{L = \sum_{ij}^{N} A_{ij}}$$

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

Weighted networks: $A_{ij} = w_{ij}$

Max number of links

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

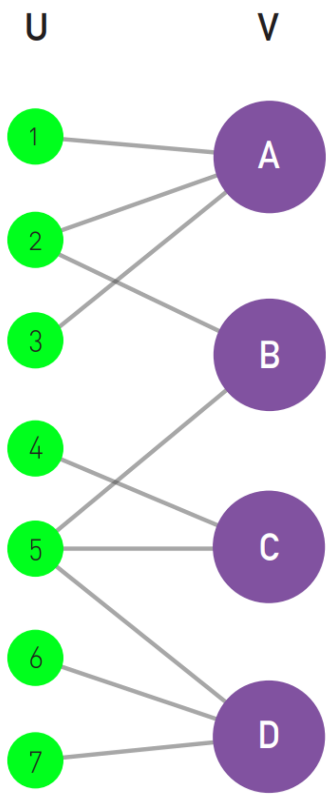


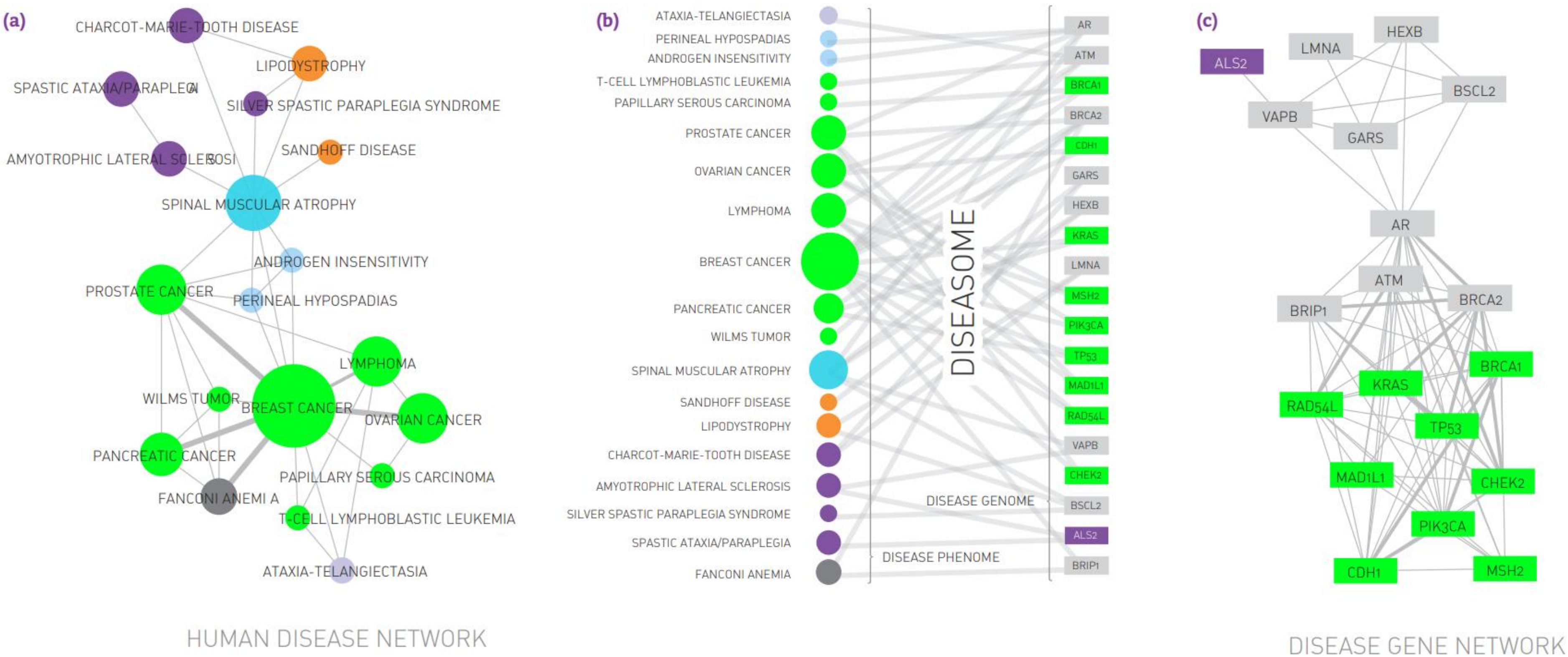
Real networks are sparse: only a fraction of the possible links are present

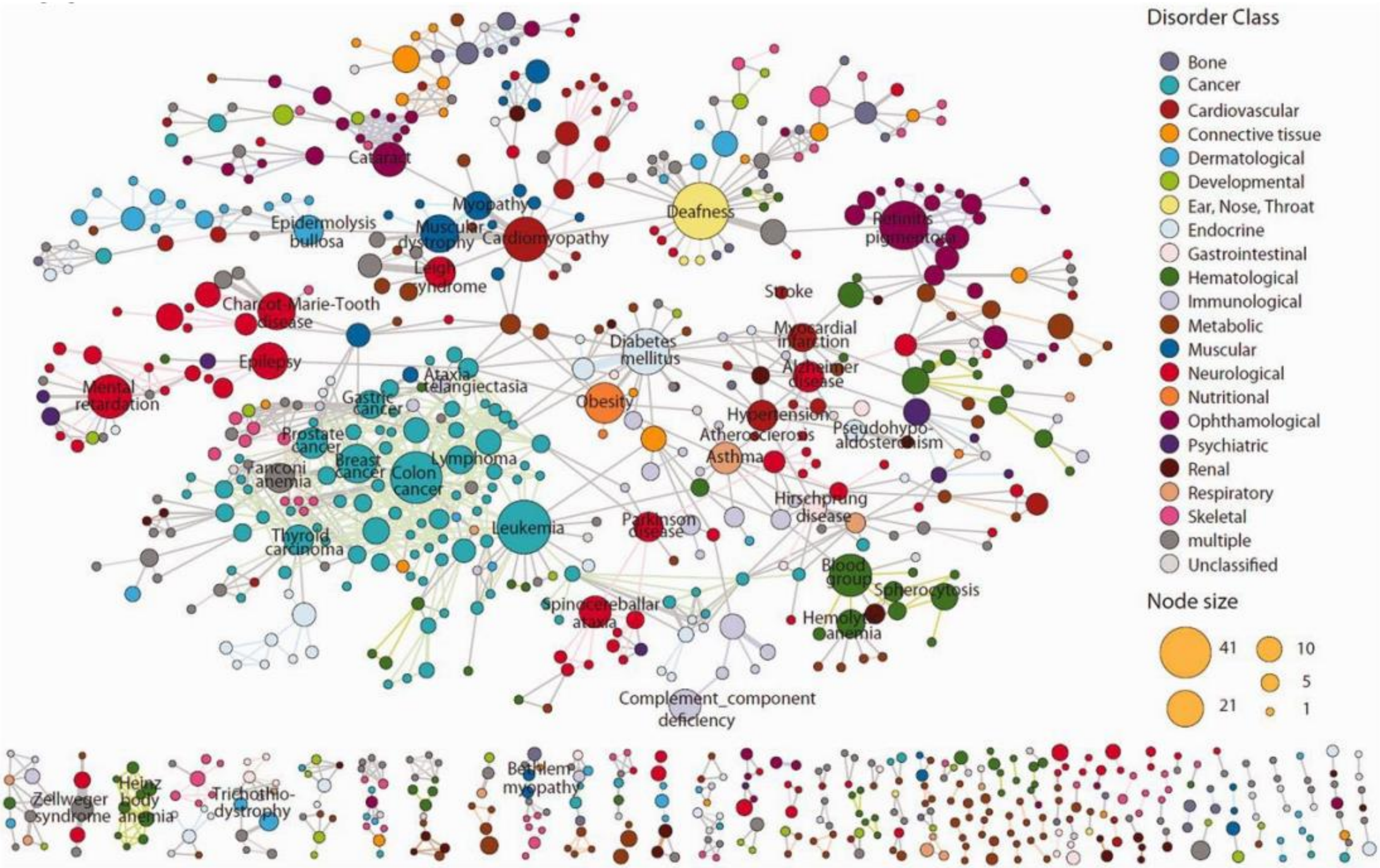
Bipartite networks

A network whose nodes are divided into 2 disjoint sets U and V such that each link connects a U-node to a V-node.

2 projections




(a)
CHARCOT-MARIE-TOOTH DISEASE
LIPODYSTROPHY
SPASTIC ATAXIA/PARAPLEGIA
SILVER SPASTIC PARAPLEGIA SYNDROME
AMYOTROPHIC LATERAL SCLEROSIS
SANDHOFF DISEASE
SPINAL MUSCULAR ATROPHY
ANDROGEN INSENSITIVITY
PROSTATE CANCER
PERINEAL HYPOSPADIAS
LYMPHOMA
WILMS TUMOR
BREAST CANCER
OVARIAN CANCER
PANCREATIC CANCER
PAPILLARY SEROUS CARCINOMA
FANCONI ANEMIA
T-CELL LYMPHOBLASTIC LEUKEMIA
ATAXIA-TELANGIECTASIA
HUMAN DISEASE NETWORK
(b)
ATAXIA-TELANGIECTASIA
PERINEAL HYPOSPADIAS
ANDROGEN INSENSITIVITY
T-CELL LYMPHOBLASTIC LEUKEMIA
PAPILLARY SEROUS CARCINOMA
PROSTATE CANCER
OVARIAN CANCER
LYMPHOMA
BREAST CANCER
PANCREATIC CANCER
WILMS TUMOR
SPINAL MUSCULAR ATROPHY
SANDHOFF DISEASE
LIPODYSTROPHY
CHARCOT-MARIE-TOOTH DISEASE
AMYOTROPHIC LATERAL SCLEROSIS
SILVER SPASTIC PARAPLEGIA SYNDROME
SPASTIC ATAXIA/PARAPLEGIA
FANCONI ANEMIA
DISEASOME
DISEASE GENOME
DISEASE PHENOME
AR
ATM
BRCA1
BRCA2
CDH1
GARS
HEXB
KRAS
LMNA
MSH2
PIK3CA
TP53
MAD1L1
RAD54L
VAPB
CHEK2
BSCL2
ALS2
BRIP1
(c)
HEXB
LMNA
ALS2
BSCL2
VAPB
GARS
AR
ATM
BRIP1
BRCA2
BRCA1
KRAS
RAD54L
TP53
MAD1L1
CHEK2
PIK3CA
CDH1
MSH2
DISEASE GENE NETWORK

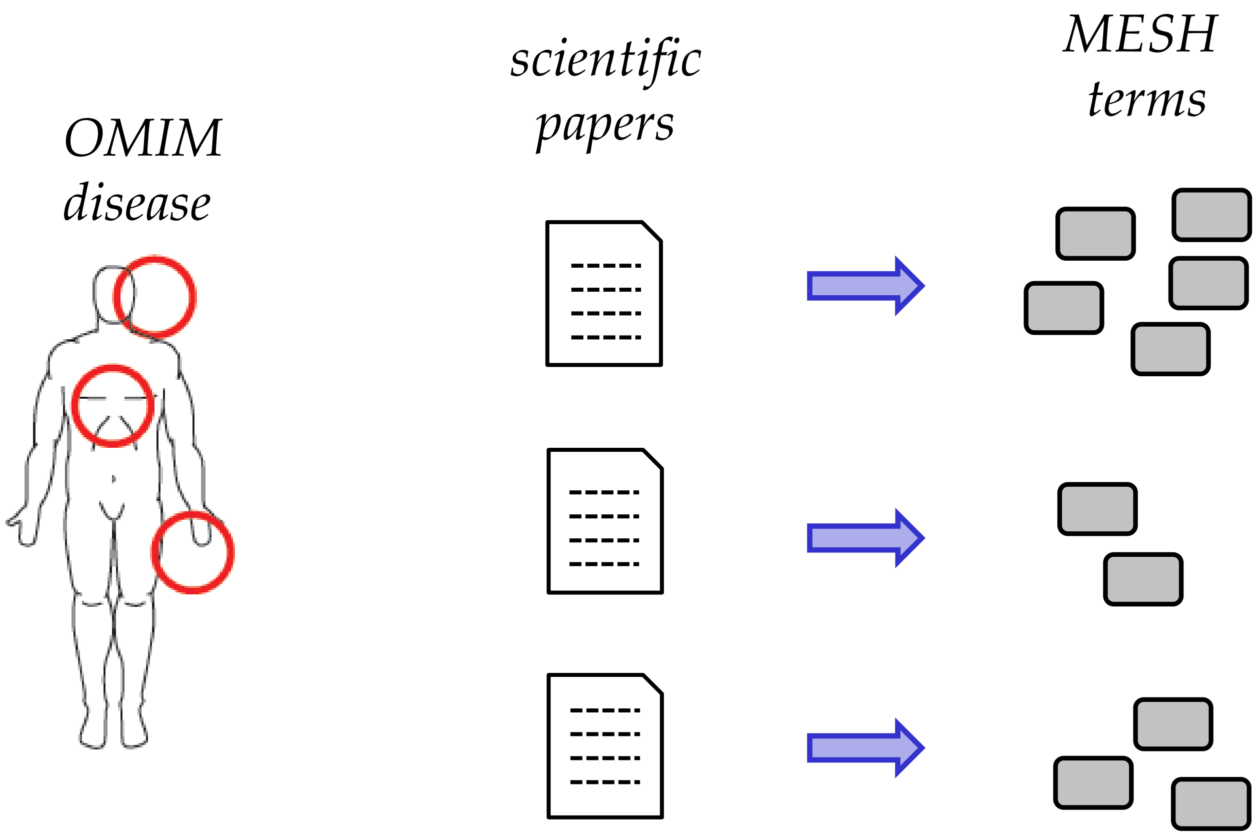

Disorder Class
Bone
Cancer
Cardiovascular
Connective tissue
Dermatological
Developmental
Ear, Nose, Throat
Endocrine
Gastrointestinal
Hematological
Immunological
Metabolic
Muscular
Neurological
Nutritional
Ophthamological
Psychiatric
Renal
Respiratory
Skeletal
multiple
Unclassified
Node size
41
10
5
21
1
Cataract
Myopathy
Deafness
Retinitis pigmentosa
Epidermolysis bullosa
Muscular dystrophy
Cardiomyopathy
Leigh syndrome
Charcot-Marie-Tooth disease
Stroke
Diabetes mellitus
Myocardial infarction
Epilepsy
Mental retardation
Ataxia telangiectasia
Alzheimer disease
Gastric cancer
Obesity
Hypertension
Prostate cancer
Atherosclerosis
Pseudohypo-aldosteronism
Breast cancer
Lymphoma
Asthma
Fanconi anemia
Colon cancer
Hirschprung disease
Thyroid carcinoma
Leukemia
Parkinson disease
Blood group
Spherocytosis
Spinocereballar ataxia
Hemolytic anemia
Complement_component deficiency
Zellweger syndrome
Heinz body anemia
Trichothio-dystrophy
Bethlem myopathy

©

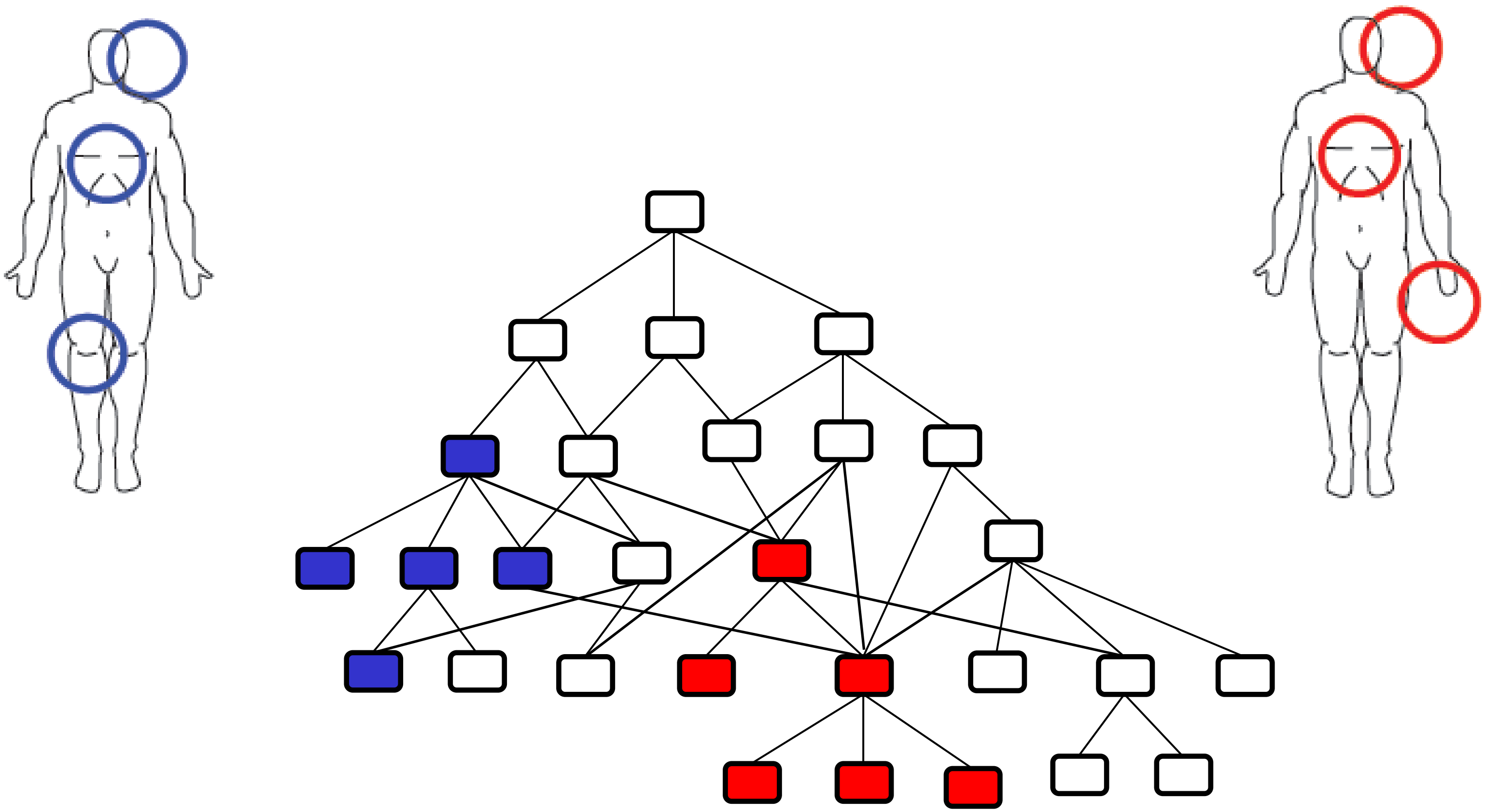
Defining a distance between diseases

[Caniza, Romero, Paccanaro, *Nature Scientific Reports*, 2015]

STEP 1: Translate a genetic disease into a set of MeSH terms



STEP 2: quantify a distance between two sets of terms on an ontology

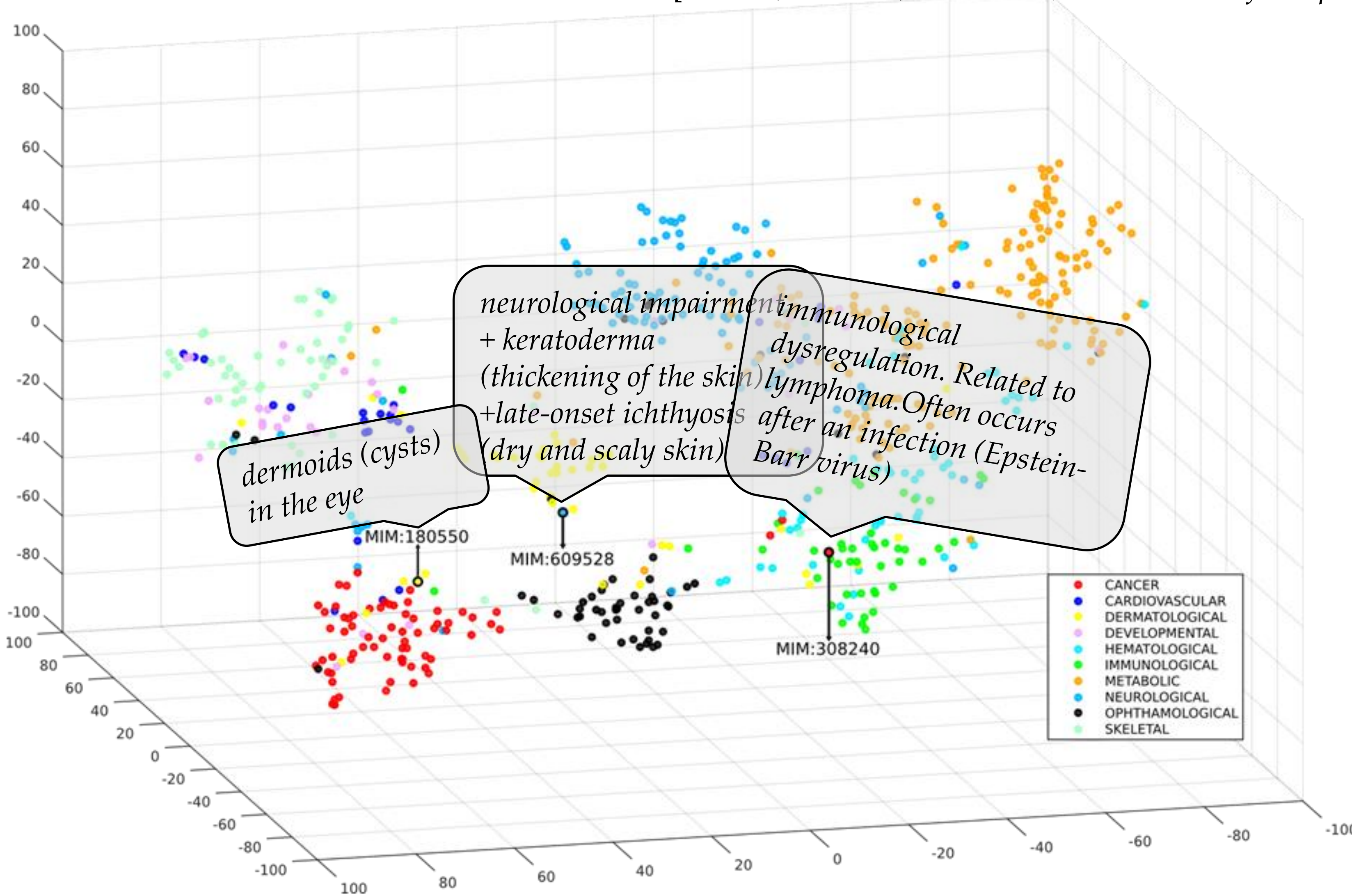


Luckily ☺ , we had developed a measure for that !

(Yang et al, *Bioinformatics*, 2012; Caniza et al, *Bioinformatics*, 2014)

Embedding diseases in 3D

[Caniza, Romero, Paccanaro, *Nature Scientific Reports*, 2015]



MIM:180550 - Ring Dermoid of Cornea – cancer/dermatological/ophthalmological

MIM:609528 - Cerebral dysgenesis, neuropathy, ichthyosis, and palmoplantar keratoderma syndrome – neurol./dermatol.

MIM:308240 - Lymphoproliferative syndrome – cancer/immunological

Paths & distances

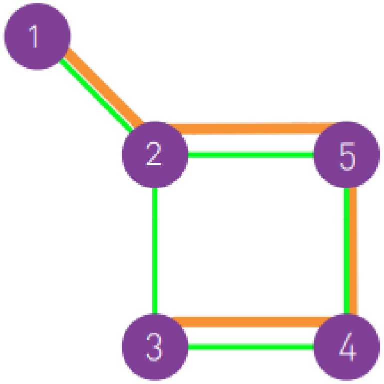
Path: a route along the links of the network.

Path length: the number of links in the path.

Distance between two nodes: length of the **shortest path** between the nodes

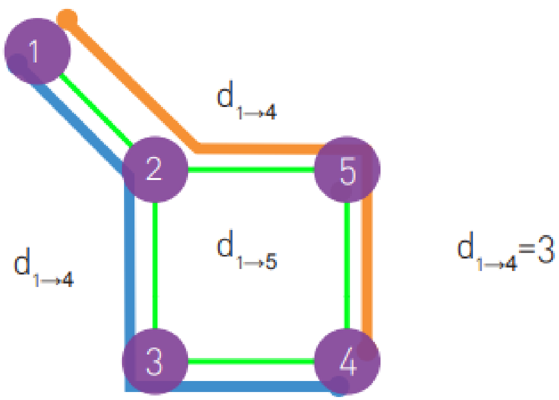
Undirected network: $d_{ij} = d_{ji}$, always

Directed network: $d_{ij} \neq d_{ji}$, in general



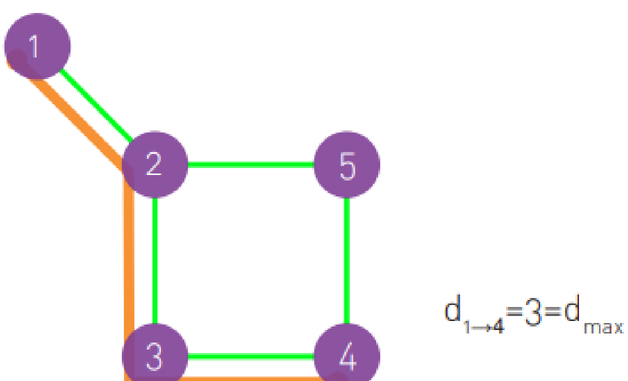
Path

A sequence of nodes such that each node is connected to the next node along the path by a link. Each path consists of <u>n+1 nodes and n links</u>. The length of a path is the number of its links, counting multiple links multiple times. For example, the orange line 1 → 2 → 5 → 4 → 3 covers a path of length four.



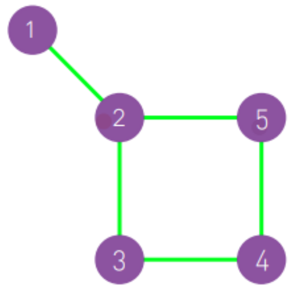
Shortest Path (Geodesic Path, d)

The path with the shortest distance d between two nodes. We also call d the distance between two nodes. Note that <u>the shortest path does not need to be unique</u>: between nodes 1 and 4 we have two shortest paths, 1→ 2→ 3→ 4 (blue) and 1→ 2→ 5→ 4 (orange), having the same length $d_{1,4}$=3.



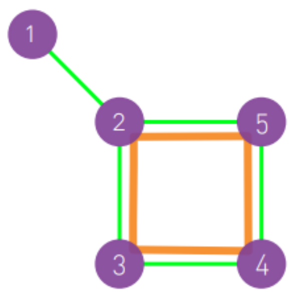
Diameter (d_{max})

<u>The longest shortest path in a graph</u>, or the distance between the two furthest nodes. In the graph shown here the diameter is between nodes 1 and 4, hence d_{max}=3.

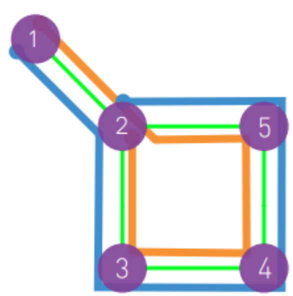


$$\langle d\rangle=[d_{1\rightarrow2}+d_{1\rightarrow3}+d_{1\rightarrow4}+d_{1\rightarrow5}+d_{2\rightarrow3}+d_{2\rightarrow4}+d_{2\rightarrow5}+d_{3\rightarrow4}+d_{3\rightarrow5}+d_{4\rightarrow5}]/10=1.6$$

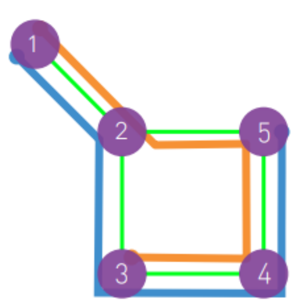
Average Path Length ($\langle d\rangle$)
The average of the shortest paths between all pairs of nodes. For the graph shown on the left we have $\langle d\rangle$=1.6, whose calculation is shown next to the figure.



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.



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

Number of shortest paths

Note that: the number of paths of length 2 between *i* and *j* is

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

The number of paths of length *d* between *i* and *j* is

$$N_{ij}^{(d)} = \sum_{k=1}^{N} A_{ij}^{d}$$

works for both undirected and directed networks

Distance between nodes *i* and *j* is the smallest *d* for which $N_{ij}(d) > 0$

Breadth first search (BFS) algorithm

To find the shortest path between node *i* and *j:*

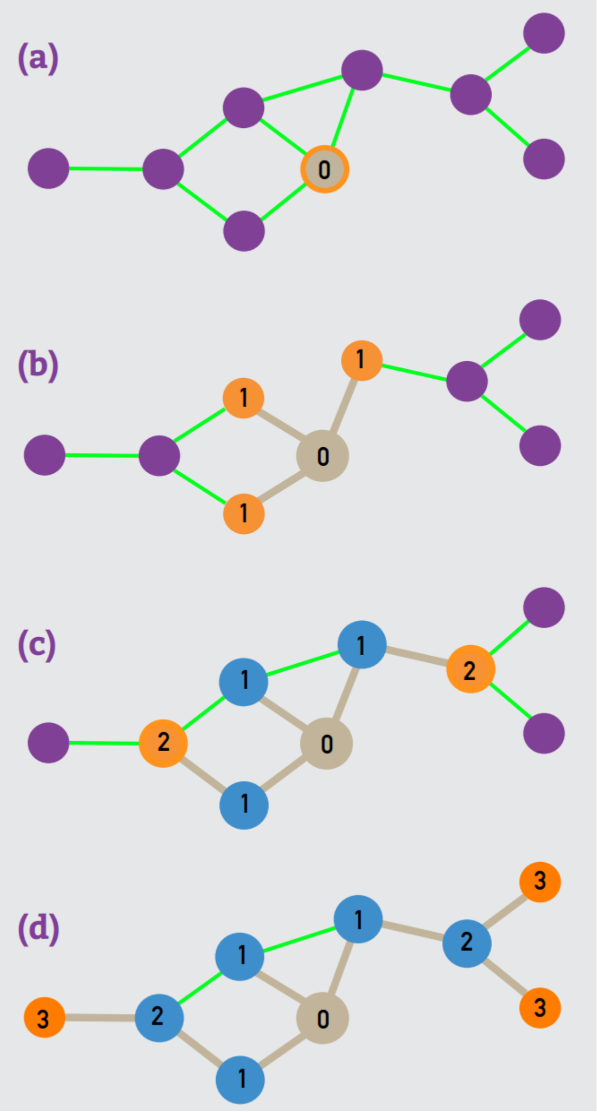
1. Start at node *i*, that we label with “0”.

2. Find the nodes directly linked to *i*. Label them distance “1” and put them in a queue.

3. Take the first node, 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.

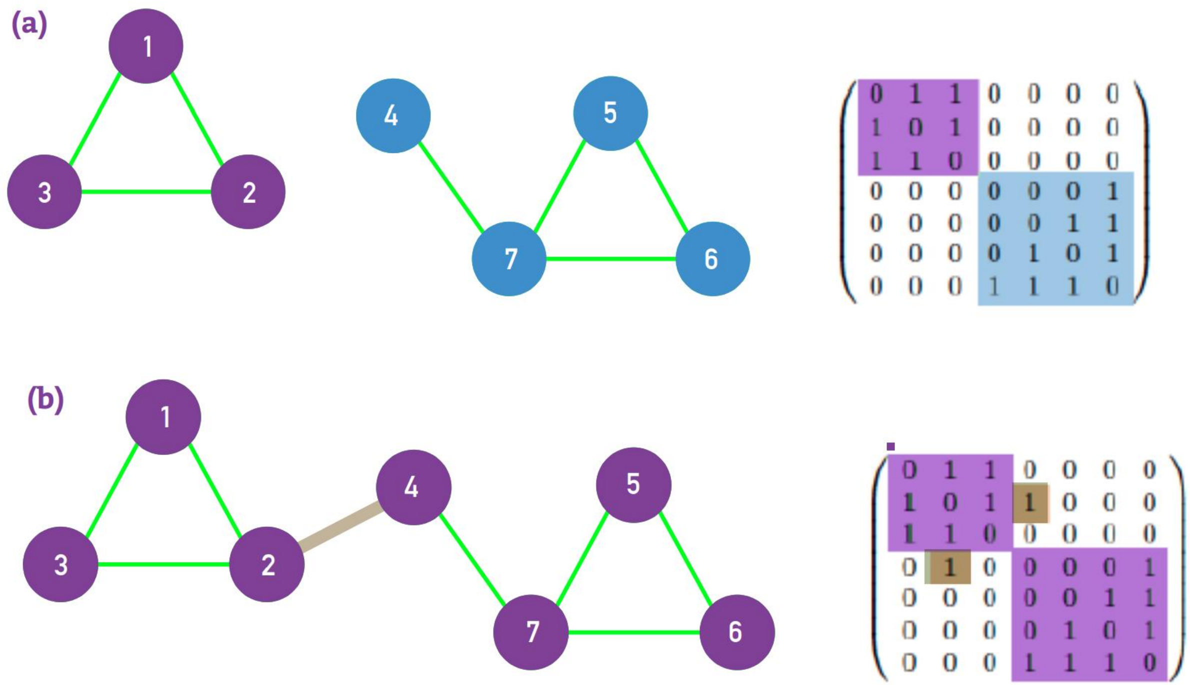
4. Repeat step 3 until you find the target node *j* or there are no more nodes in the queue.

5. The distance between *i* and *j* is the label of *j*. If *j* does not have a label, then $d_{ij} = \infty$.



Connectedness

- Connected/disconnected components (or clusters)
- Bridges



Components are efficiently identified using BFS

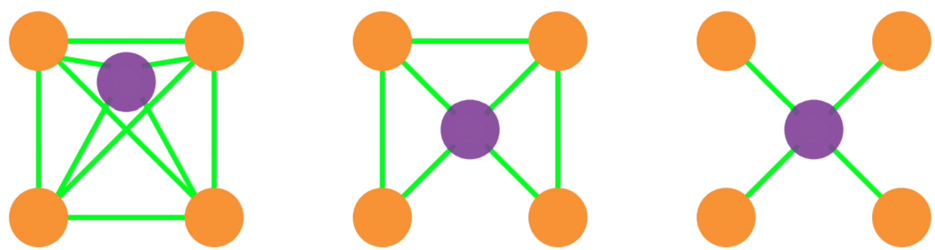
Clustering Coefficient

It captures the degree to which the neighbours of a node link to each other

Clustering coefficient for node i with degree k_i , where there are L_i links between its neighbours:

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

probability that two neighbours of a node are linked



C_i=1 $\quad$ C_i=1/2 $\quad$ C_i=0

Average clustering coefficient:

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

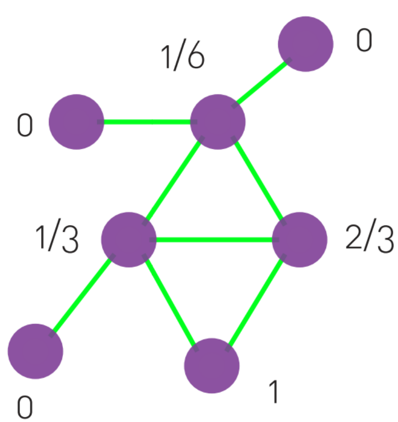
probability that two neighbours of a randomly selected node are linked

Global clustering coefficient
(aka ratio of transitive triplets)

L_i is the **number of triangles** that a node *i* participates in.

Connected triplet is an ordered set of three nodes ABC such that A connects to B and B connects to C.

Global clustering coefficient: $$C_\Delta = \frac{3 \times Number\ Of\ Triangles}{Number\ Of\ Connected\ Triples}$$

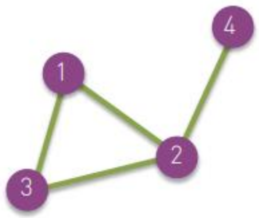


$$\langle C \rangle = \frac{13}{42} \approx 0.310$$

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

A summary of the most common network types

Undirected



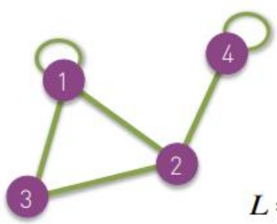
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Undirected Network
A network whose links do not have a defined direction.
Examples: Internet, power grid, science collaboration networks.

Self-loops



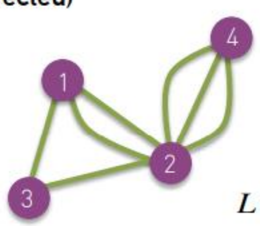
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

Self-loops
In many networks nodes do not interact with themselves, so the diagonal elements of the adjacency matrix are zero, $A_{ii} = 0$, $i = 1,\ldots, N$. In some systems self-interactions are allowed; in such networks, self-loops represent the fact that node i interacts with itself.
Examples: WWW, protein interactions.

Multigraph
(undirected)



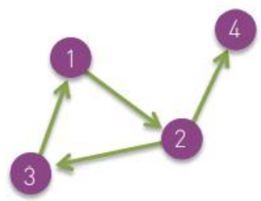
$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

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

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

Multigraph/Simple Graphs
In a multigraph nodes are permitted to have multiple links (or parallel links) between them. Hence A_{ij} can be any positive integer. Networks that do not allow multiple links are called *simple*.
Multigraph Examples: Social networks, where we distinguish friendship, family and professional ties.

Directed



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

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

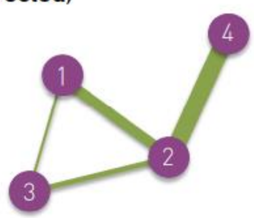
$$L = \sum_{i,j=1}^{N} A_{ij} \qquad <k>= \frac{L}{N}$$

Directed Network

A network whose links have selected directions. Examples: WWW, mobile phone calls, citation network.

Weighted

(undirected)



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

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

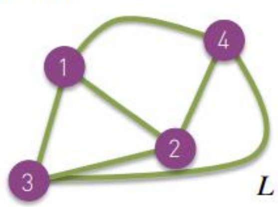
$$<k>= \frac{2L}{N}$$

Weighted Network

A network whose links have a defined weight, strength or flow parameter. The elements of the adjacency matrix are $A_{ij} = w_{ij}$ if there is a link with weight w_{ij} between them. For unweighted (binary) networks, the adjacency matrix only indicates the presence ($A_{ij} = 1$) or the absence ($A_{ij} = 0$) of a link.
Examples: Mobile phone calls, email network.

Complete Graph

(undirected)



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

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

$$L = L_{\max} = \frac{N(N-1)}{2} \qquad <k>= N-1$$

Complete Graph (Clique)

In a complete graph, or a clique, all nodes are connected to each other.
Examples: Actors in the cast of the same movie, as they are all linked to each other in the actor network.