Chapter 2

Networks & Graphs: Basic Measures

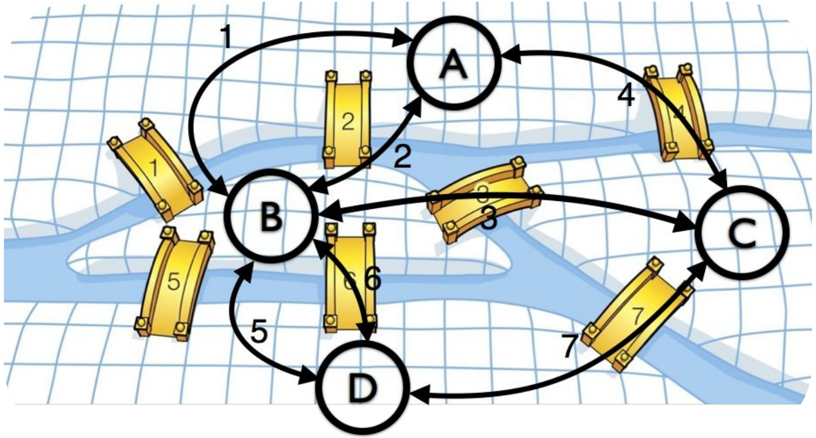
Summary
- Graph representations
- Type of Networks
- Degree distribution
- Paths & Connectedness
- Clustering

Reading
- Chapter 2 of Barabasi's book

The Bridges of Konigsberg

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

Famous Konigsberg Citizens
Immanuel Kant (philosopher, 1724-1804)



Euler’s theorem (1735)

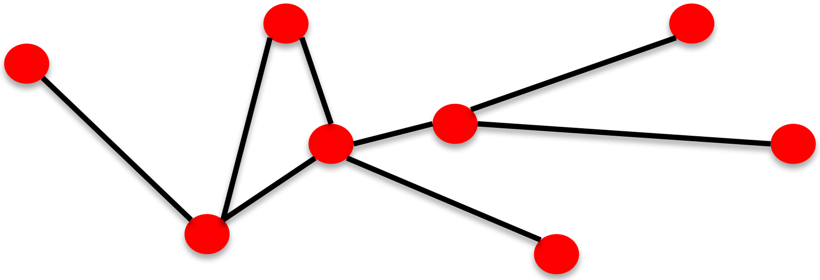
- If a graph has more than two nodes of odd degree, there is no path/cycle that crosses each bridge exactly once.
- If a graph is connected and has no odd degree nodes, it has at least one path.

Components of a Complex System

Networks or Graphs?

Network *<nodes, links>*
refers to real systems
(www, social network, metabolic network)

Graph *<vertices, edges>*
mathematical representation of a network
(web graph, social graph)



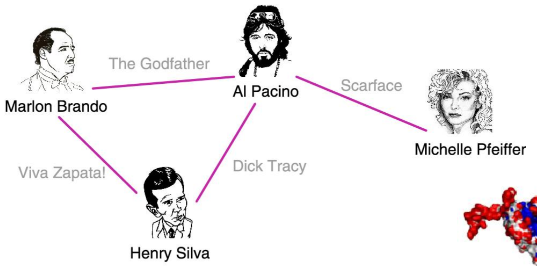
		Symbol
Components	nodes, vertices	N
Interactions	edges, links	L
System	network, graph	(N,L)

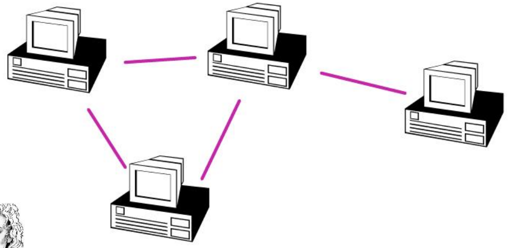
A Common Language

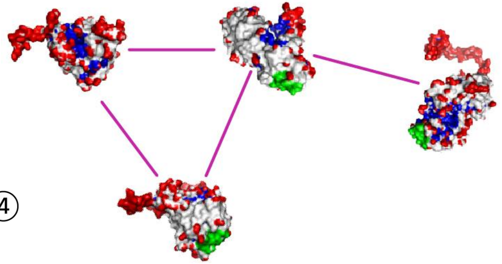
The choice of the proper network representation determines our ability to use network theory successfully.

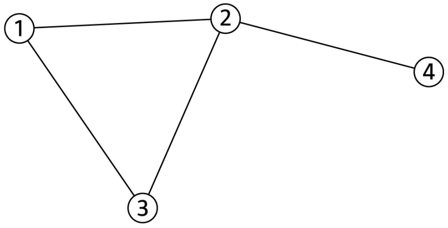
In some cases there is a unique, unambiguous representation. In other cases, the representation is by no means unique.

The way we assign the links between a group of individuals will determine the nature of the question we can study.





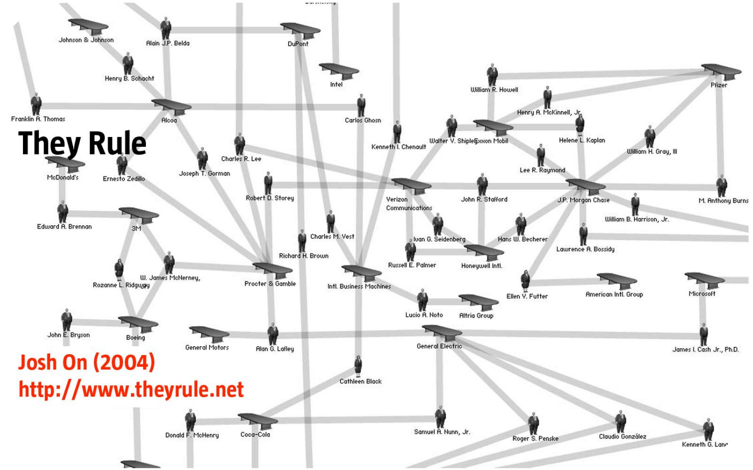




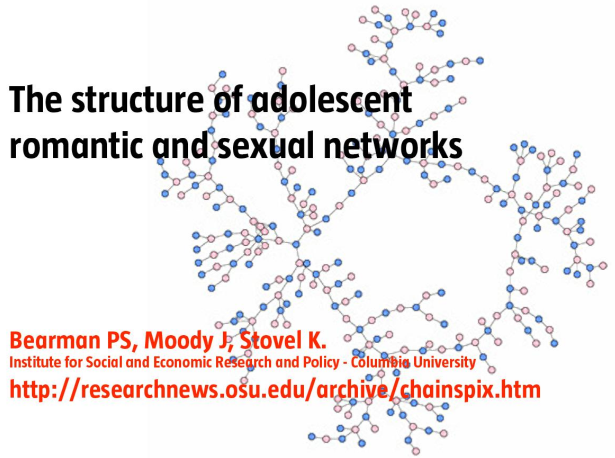
N=4 L=4

Proper representations (examples)

If you connect individuals that work with each other, you will explore the *professional network*.



If you connect those that have a romantic and sexual relationship, you will be exploring the *sexual networks*.



If you connect individuals based on their first name

(e.g., *all Peters connected to each other*),

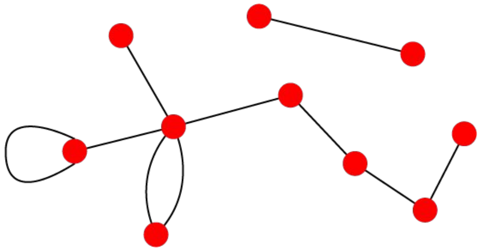
you will be exploring what?

It is a network, *nevertheless*.

Directedness

Undirected graphs

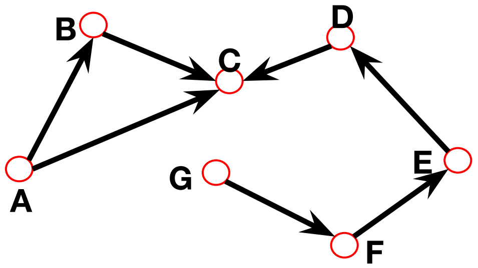
Links: undirected (*symmetrical*)



Examples of Undirected links
- Co-authorship links
- Actor network
- Protein interactions

Directed graphs (DiGraphs)

Links: directed (*arcs*).



Example of Directed links
- URLs on the www
- Phone calls
- Metabolic reactions

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L
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

Reference Networks

Degree,
Average Degree,
Degree Distribution

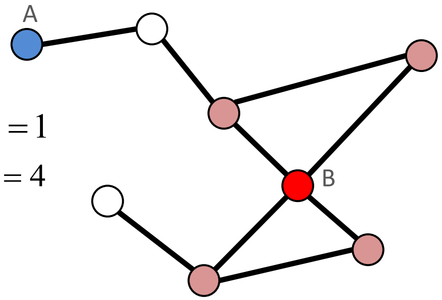
Node Degree

Undirected graphs

the number of links connected to the node

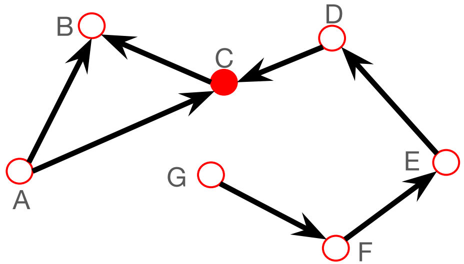
$k_A = 1$

$k_B = 4$



Directed graphs (DiGraphs)

we can define an in-degree and out-degree.
The (total) degree is the sum of in- and out-degree.



$k_C^{in} = 2$

$k_C^{out} = 1$

$k_C = 3$

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

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

A Bit of Statistics

Four key quantities characterize a sample of N values $x_1,...,x_n$

Average (mean):

$$\langle x \rangle = \frac{x_1 + x_2 + \dots .1x_3}{N} = \frac{1}{N}\sum_{i=1}^{n} x_i$$

The n^{th} moment:

$$\langle x^a \rangle = \frac{x_1^n + x_2^n + \dots - x_n^n}{N} = \frac{1}{N}\sum_{k=1}^{N} x_i^n$$

The Standard deviation:

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

The Distribution of *x*:

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

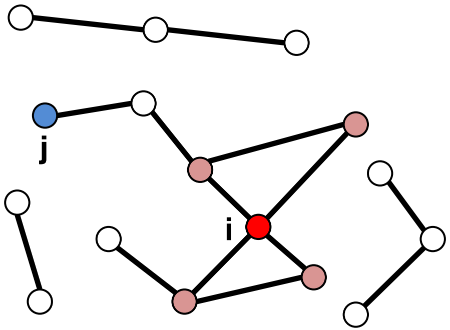
where p_x follows

$$\sum_x p_x = 1\left(\int p_x dx = 1\right)$$

Average Degree

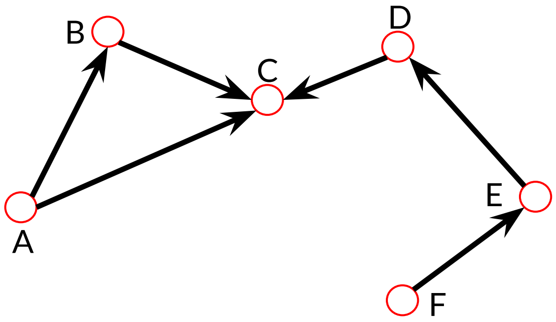
Undirected graphs

N - the number of nodes in the graph



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

Directed graphs (DiGraphs)



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

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

NETWORK	NODES	LINKS	DIRECTED UNDIRECTED	N	L	⟨k⟩
Internet	Routers	Internet connections	Undirected	192,244	609,066	6.33
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

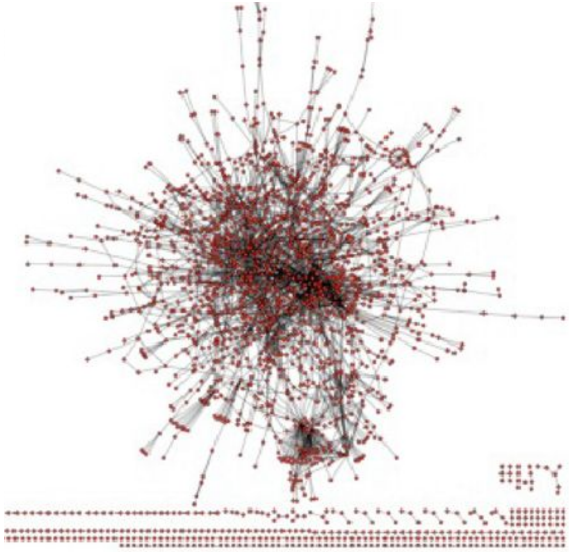
Reference Networks: Average Degree

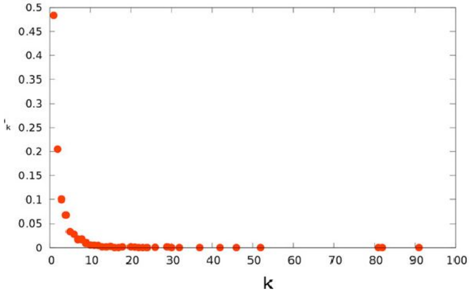
Degree Distribution

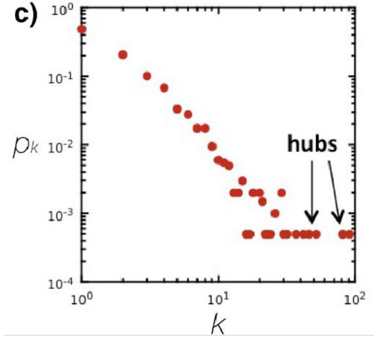
P(k): probability that a randomly chosen node has degree *k*

N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot







Degree Distribution (cont'd)

Discrete Representation: p_k is the probability that a node has degree k.

Continuum Description: p(k) is the pdf of the degrees, where

$$\int_{k_1}^{k} p(k)dk$$

represents the probability that a node's degree is between k_1 and k_2.

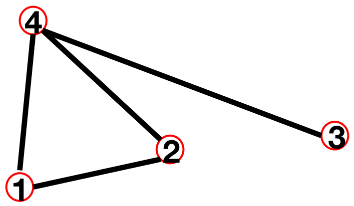
Normalization condition:

$$\sum_{a}^{\infty} p_k = 1 \qquad \int_{\kappa_{min}} p(k)dk = 1$$

where K_{min} is the minimal degree in the network.

Adjacency matrix

Undirected graphs



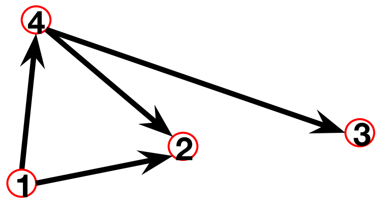
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$k_i = \sum_{j=1}^{N} A_{ij} \qquad k_j = \sum_{i=1}^{N} A_{ij}$$

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

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

Directed graphs (DiGraphs)



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

$$k_i^{in} = \sum_{j=1}^{N} A_{ij} \qquad k_j^{out} = \sum_{i=1}^{N} A_{ij}$$

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

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

Paths and Connectedness

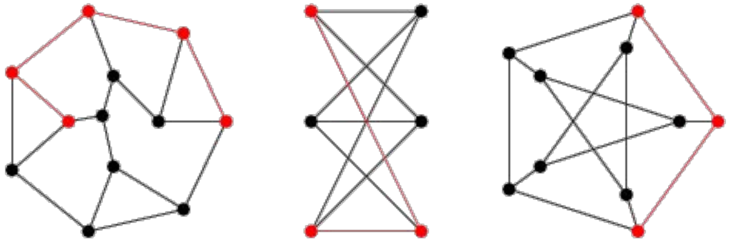
Paths

A *path is* a sequence of nodes in which each node is adjacent to the next one

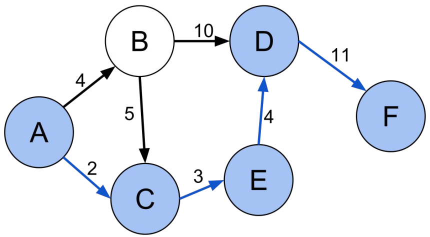
$P_{i0,in}$ of length n between nodes i_0 and i_n is an ordered collection of $n+1$ nodes and n links

$$P_n = \{i_0, i_1, i_2, \ldots, i_n\}$$

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



Examples of phats in an undirected graph.

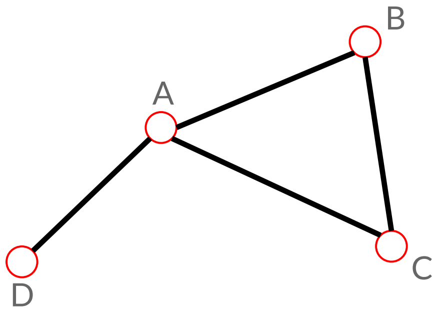


In a directed graph, the path can follow only the direction of an arrow.

Distance in a Graph

Undirected graphs

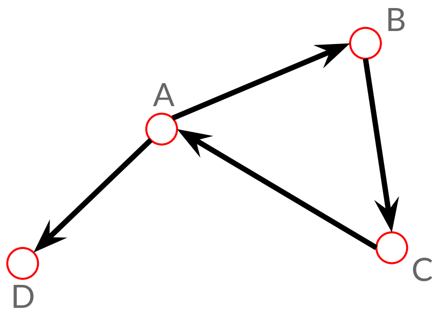
The *distance* (*shortest path*, *geodesic path*) between two nodes is defined as the number of edges along the shortest path connecting them.



*If the two nodes are disconnected, the distance is infinity.

Directed graphs (DiGraphs)

Each path needs to follow the direction of the arrows.



Thus in a digraph the distance from node A to B (on an AB path) is generally different from the distance from node B to A (on a BCA path).

Number of paths between two nodes

N_{ij}, number of paths between any two nodes ***i*** and ***j***:

Length n=1:
If there is a link between *i* and *j*, then $A_{ij}=1$ and $A_{ij}=0$ otherwise.

Length n=2:
If there is a path of length two between *i* and *j*, then $A_{ik}A_{kj}=1$, and $A_{ik}A_{kj}=0$ otherwise.

The number of paths of length 2:

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

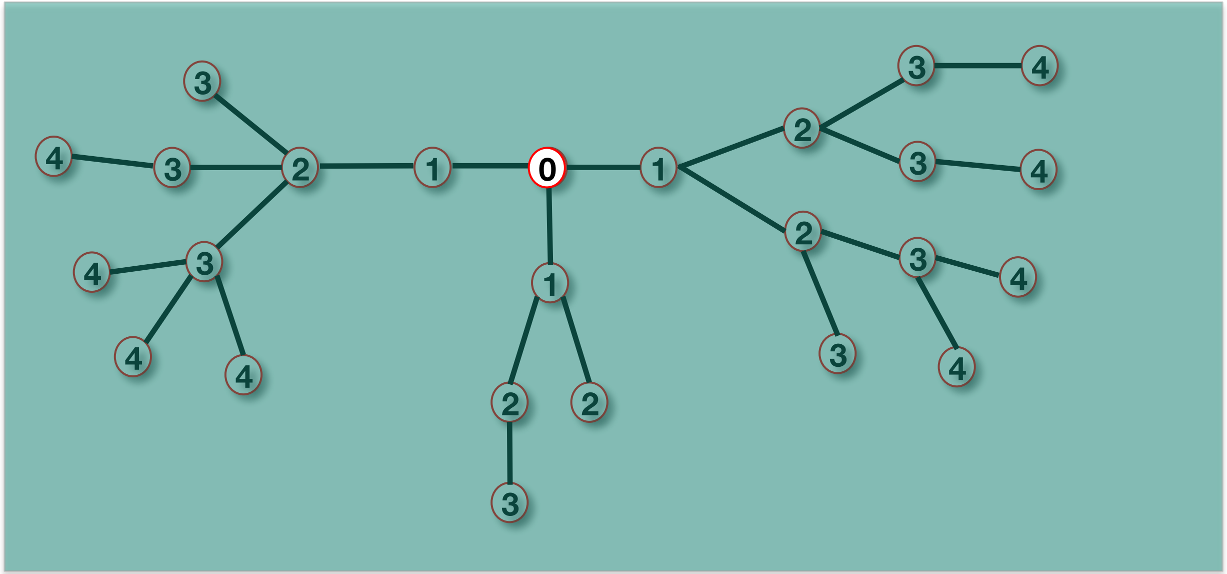
Length n:
In general, if there is a path of length *n* between *i* and *j*, then $A_{ik}\ldots A_{lj}=1$ and $A_{ik}\ldots A_{lj}=0$ otherwise.

The number of paths of length *n* between *i* and *j* is(*)

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

(*) *Holds for both directed and undirected networks.*

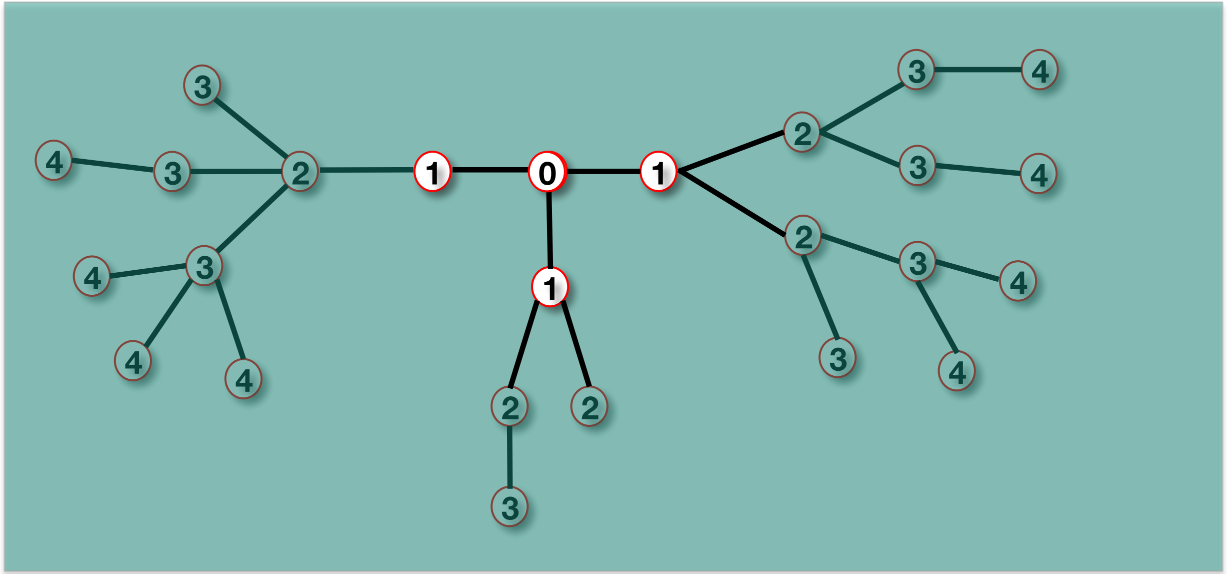
Finding Distances: BFS



Distance between node 0 and node 4:

1. Start at 0.

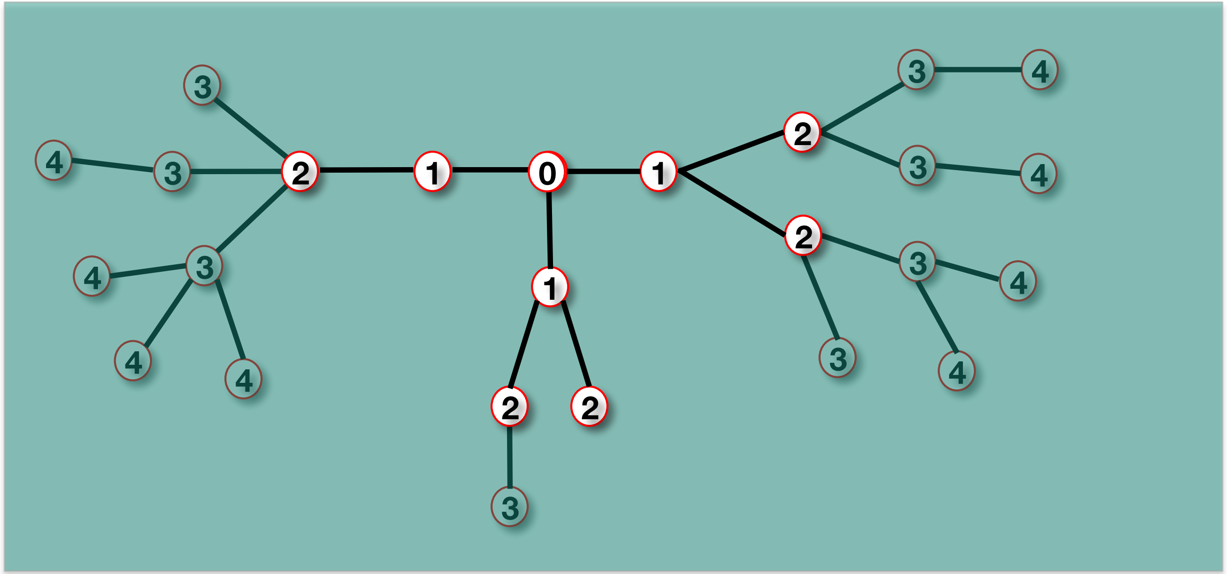
Finding Distances: BFS



Distance between node 0 and node 4:

1. Start at 0.
2. Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.

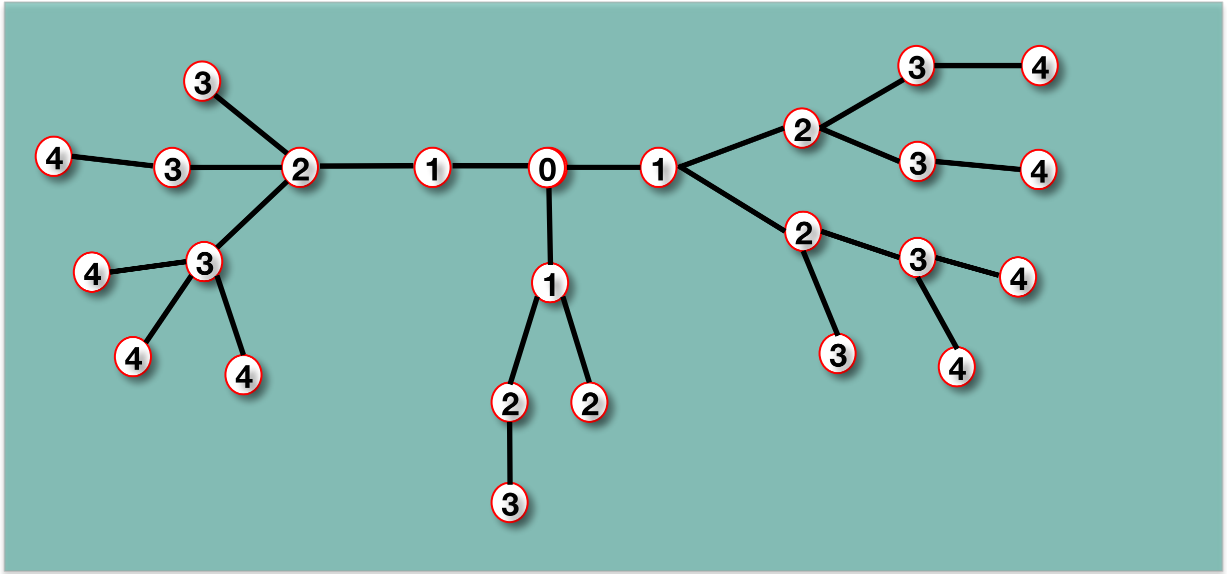
Finding Distances: BFS



Distance between node 0 and node 4:

1. Start at 0.
2. Find the nodes adjacent to 1. Mark them as at distance 1. Put them in a queue.
3. Take the first node out of the queue. Find the unmarked nodes adjacent to it in the graph. Mark them with the label of 2. Put them in the queue.

Finding Distances: BFS



Distance between node 0 and node 4:

Repeat until you find node 4 or there are no more nodes in the queue.

The distance between 0 and 4 is the label of 4 or, if 4 does not have a label, infinity.

Diameter and Average distance

Diameter (d_{max}):

the maximum distance between any pair of nodes in the graph.

Average path length/distance, <*d*>, for a connected graph:

$$\langle d \rangle \equiv \frac{1}{2L_{\max}} \sum_{i,j \neq i} d_{ij}$$

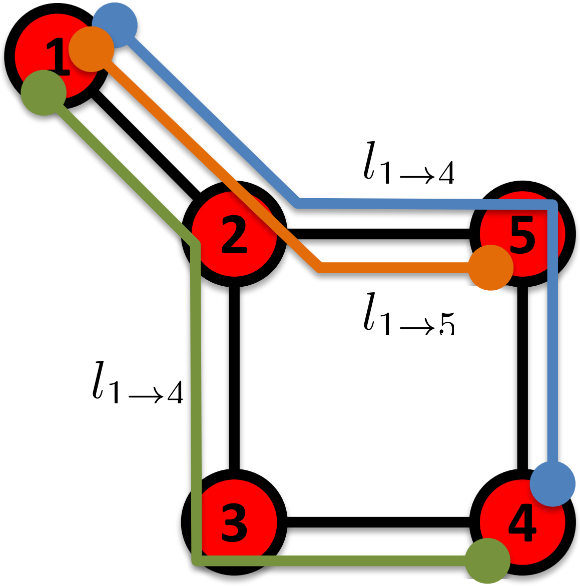
where d_{ij} is the distance from node *i* to node j

In an *undirected graph* $d_{ij} = d_{ji}$, *so* we only need to count them once:

$$\langle d \rangle \equiv \frac{1}{L_{\max}} \sum_{i,j > i} d_{ij}$$

Paths: a summary

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

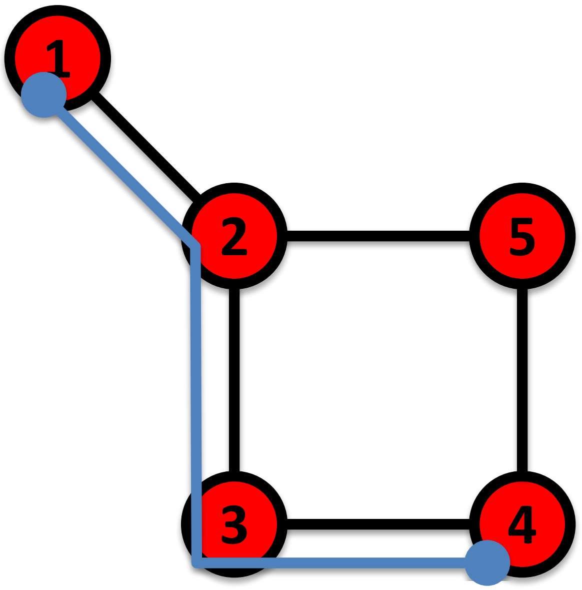


$l_{1\to4} = 3$

$l_{1\to5} = 2$

Paths: a summary

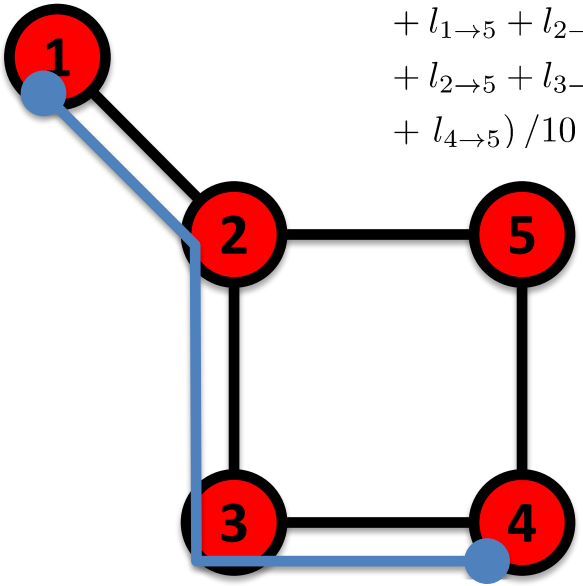
Diameter
The longest shortest path in a graph.



$$l_{1 \to 4} = 3$$

Paths: a summary

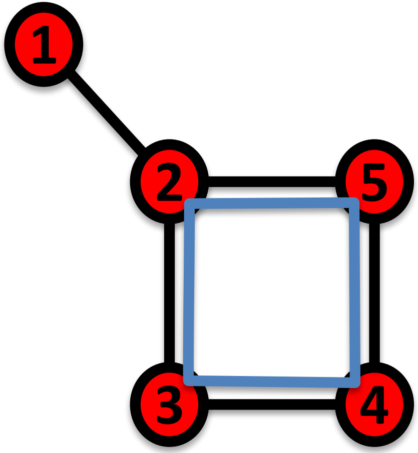
$$(l_{1\to2} + l_{1\to3} + l_{1\to4} + \\ + l_{1\to5} + l_{2\to3} + l_{2\to4} + \\ + l_{2\to5} + l_{3\to4} + l_{3\to5} + \\ + l_{4\to5})\,/10 = 1.6$$



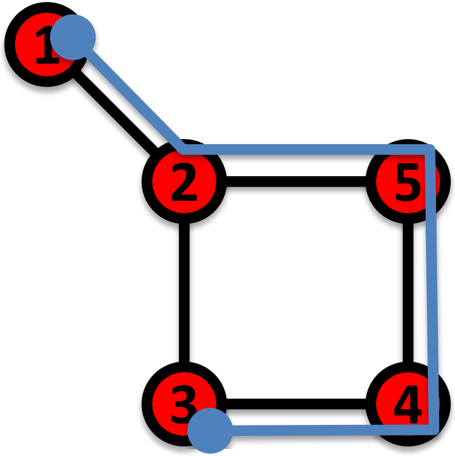
Average Path Length

The average of the shortest paths for all pairs of nodes.

Paths: a summary

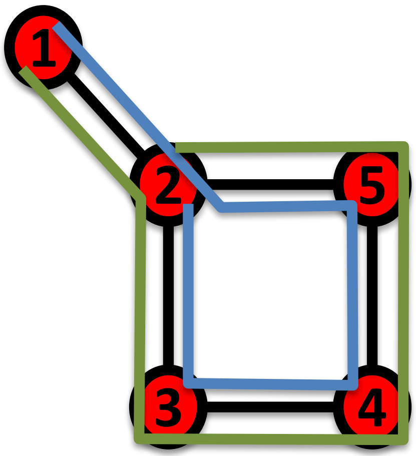


Cycle
A path with the same start and end node.

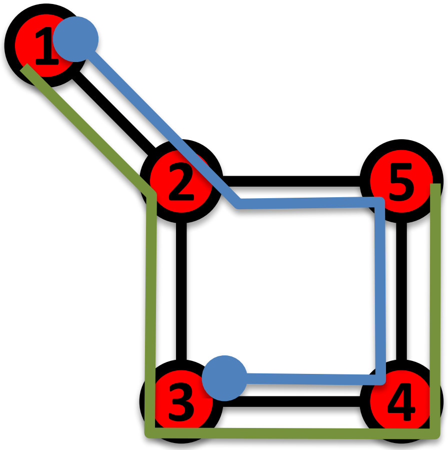


Self-Avoiding Path
A path that does not intersect itself.

Paths: a summary



Eulerian Path/Cycle
A path that traverses each link exactly once.

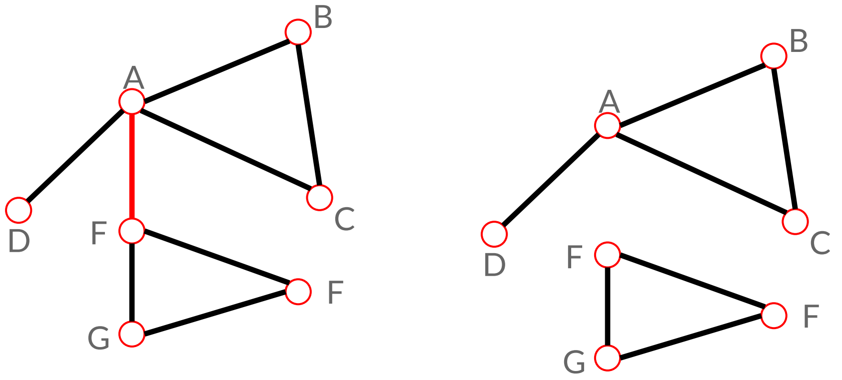


Hamiltonian Path/Cycle
A path that visits each node exactly once.

Connectivity of undirected graphs

Connected (undirected) graph:
any two vertices can be joined by a path.

A **disconnected graph** is made up by two or more connected components.

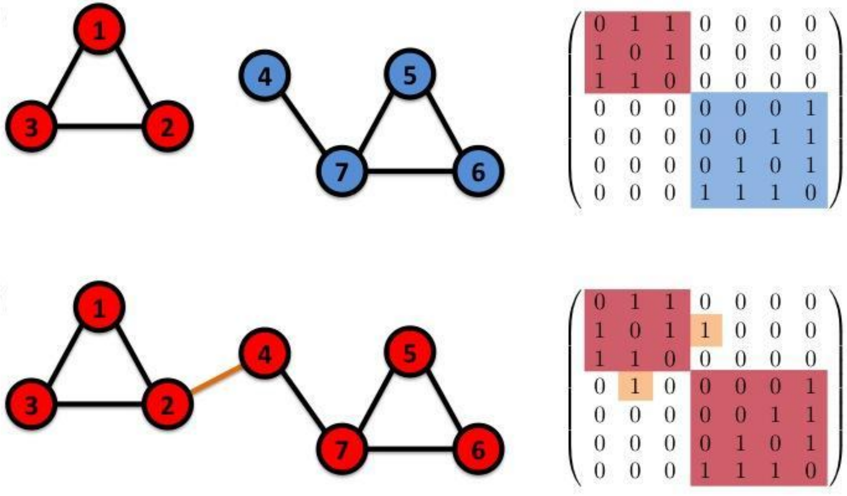


Bridge:
if we erase it, the graph becomes disconnected.
Example (A,F)

Largest Component: **Giant Component**
The rest: **Isolates**

Connectivity of undirected graphs

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.

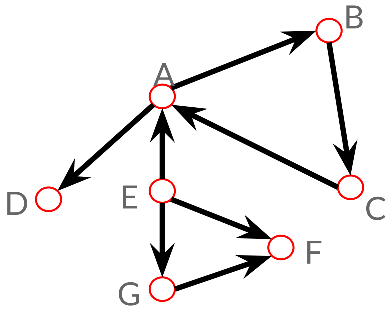


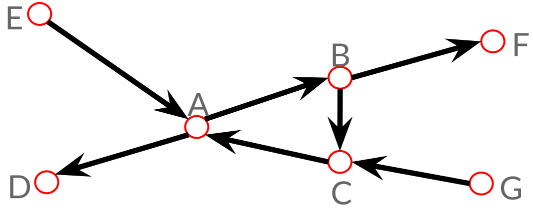
Connectivity of directed graphs

Strongly connected directed graph (SCC):
has a path from each node to every other node and vice versa (e.g. AB path and BA path).

Weakly connected directed graph (WCC):
it is connected if we disregard the edge directions.

Strongly connected components can be identified, but not every node is part of a nontrivial strongly connected component.





In-component:
nodes that can reach the scc,

Out-component:
nodes that can be reached from the scc.

Network Density

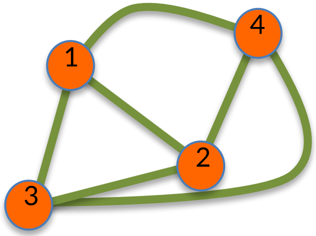
Complete Graph

A graph with degree

$$L = L_{max}$$

is called a complete graph, and its average degree is

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



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

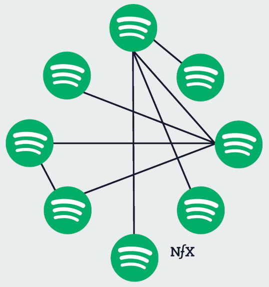
The maximum number of links an undirected network of N nodes can have is:

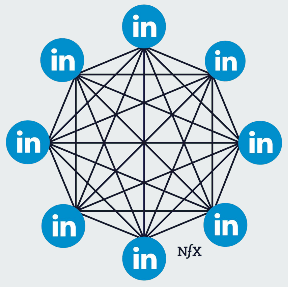
$$\boxed{L_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}}$$

What about directed networks?

Network Density

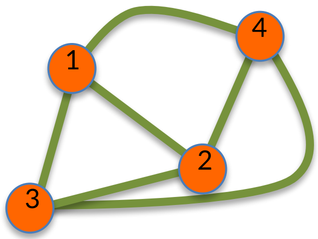
Ratio of existing edges over possible ones.



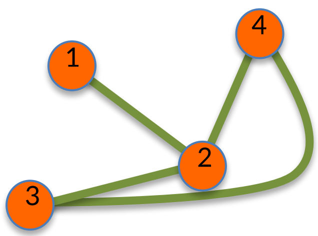


$$d(G) = \frac{L}{L_{max}}$$

Examples



Complete graph
d(G) = 1



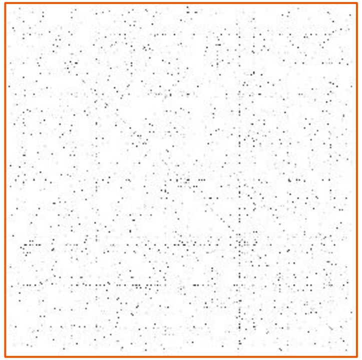
Sparse graph
d(G) = 2/3

Most networks observed in real systems are sparse

$L \ll L_{max}$

$\langle k \rangle \ll N-1$

$d(G) \ll 1$



Sparse
Adjacency matrix

WWW (ND Sample):	N=325,729;	$L=1.4\ 10^6$	$L_{max}=10^{12}$	$\langle k \rangle=4.51$
Protein (*S. Cerevisiae*):	N= 1,870;	L=4,470	$L_{max}=10^7$	$\langle k \rangle=2.39$
Coauthorship (Math):	N= 70,975;	$L=2\ 10^5$	$L_{max}=3\ 10^{10}$	$\langle k \rangle=3.9$
Movie Actors:	N=212,250;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

(Source: Albert, Barabasi, RMP2002)

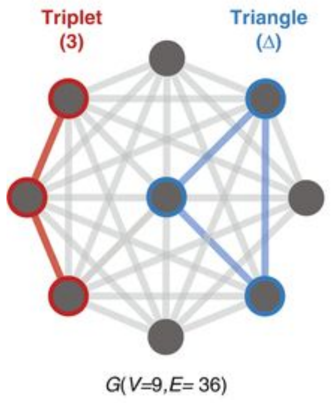
Clustering Coefficient

Clustering Coefficient

How “clustered” is my network?

Global Clustering coefficient

- Triangles and triplets
- C in [0,1]



$$C = \frac{3 \times \text{ number of triangles}}{\text{number of all triplets}}$$

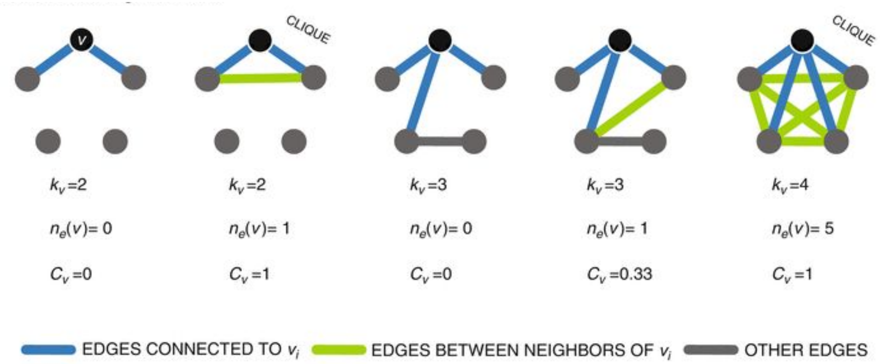
Watts & Strogatz, Nature (1998)

Clustering Coefficient

What fraction of your neighbors are connected?

Local Clustering coefficient

- Node i with degree k_i
- C_i in [0,1]



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

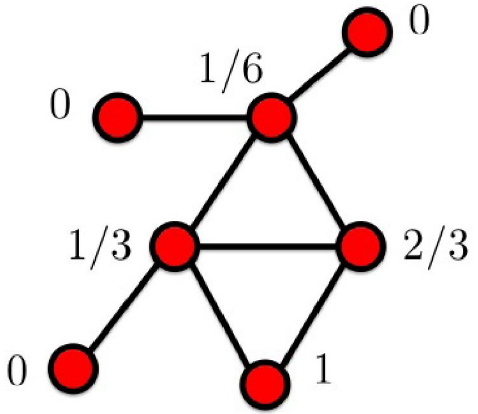
Watts & Strogatz, Nature (1998)

Clustering Coefficient

What fraction of your neighbors are connected?

Local Clustering coefficient

- Node i with degree k_i
- C_i in [0,1]



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

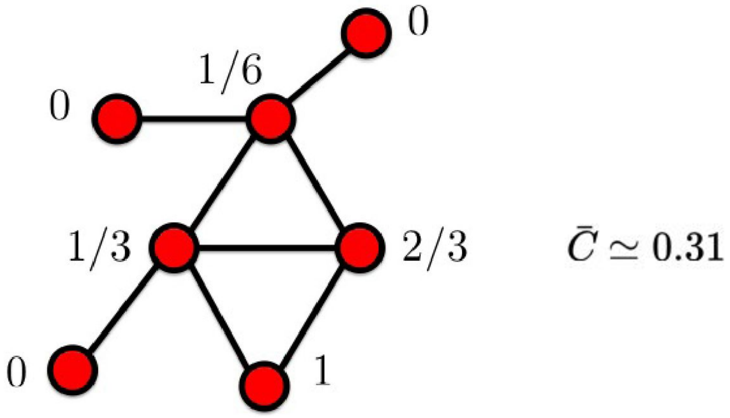
Watts & Strogatz, Nature (1998)

Clustering Coefficient

What fraction of your neighbors are connected on average?

Average Clustering coefficient

- Average of local clustering coefficients
- C in [0,1]



$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

Watts & Strogatz, Nature (1998)

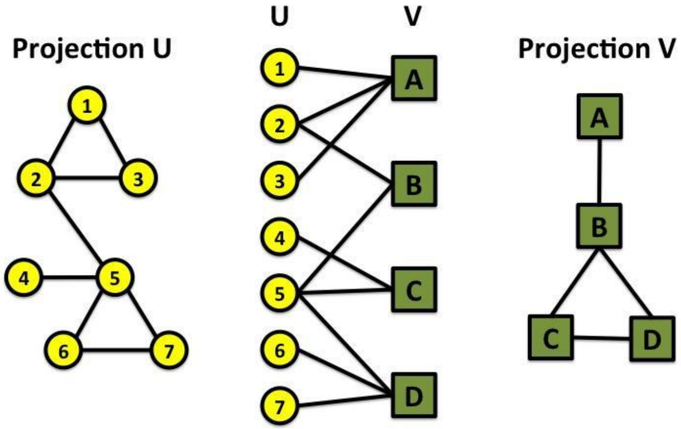
Bipartite Networks

Bipartite Graphs

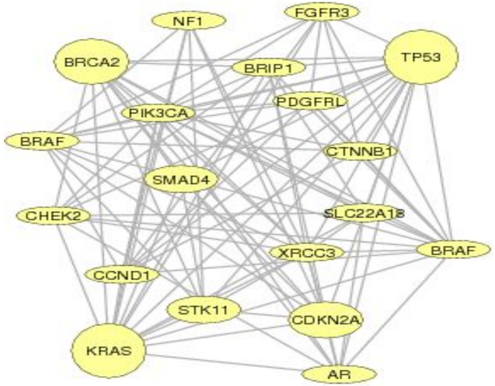
Bipartite graph (or bigraph)

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*.

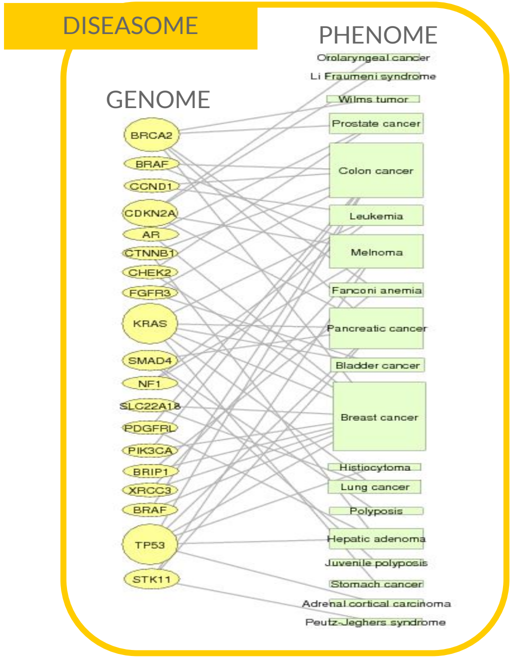
Examples
Hollywood actor network
Collaboration networks
Disease network (diseasome)

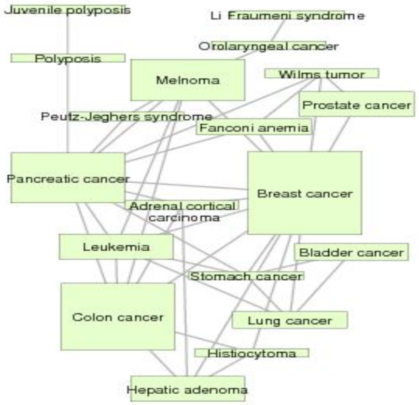


Projection
Two nodes of the **same class** are connected by a (weighted) edge if they share at least a **common neighbor**



Gene network
(projection)

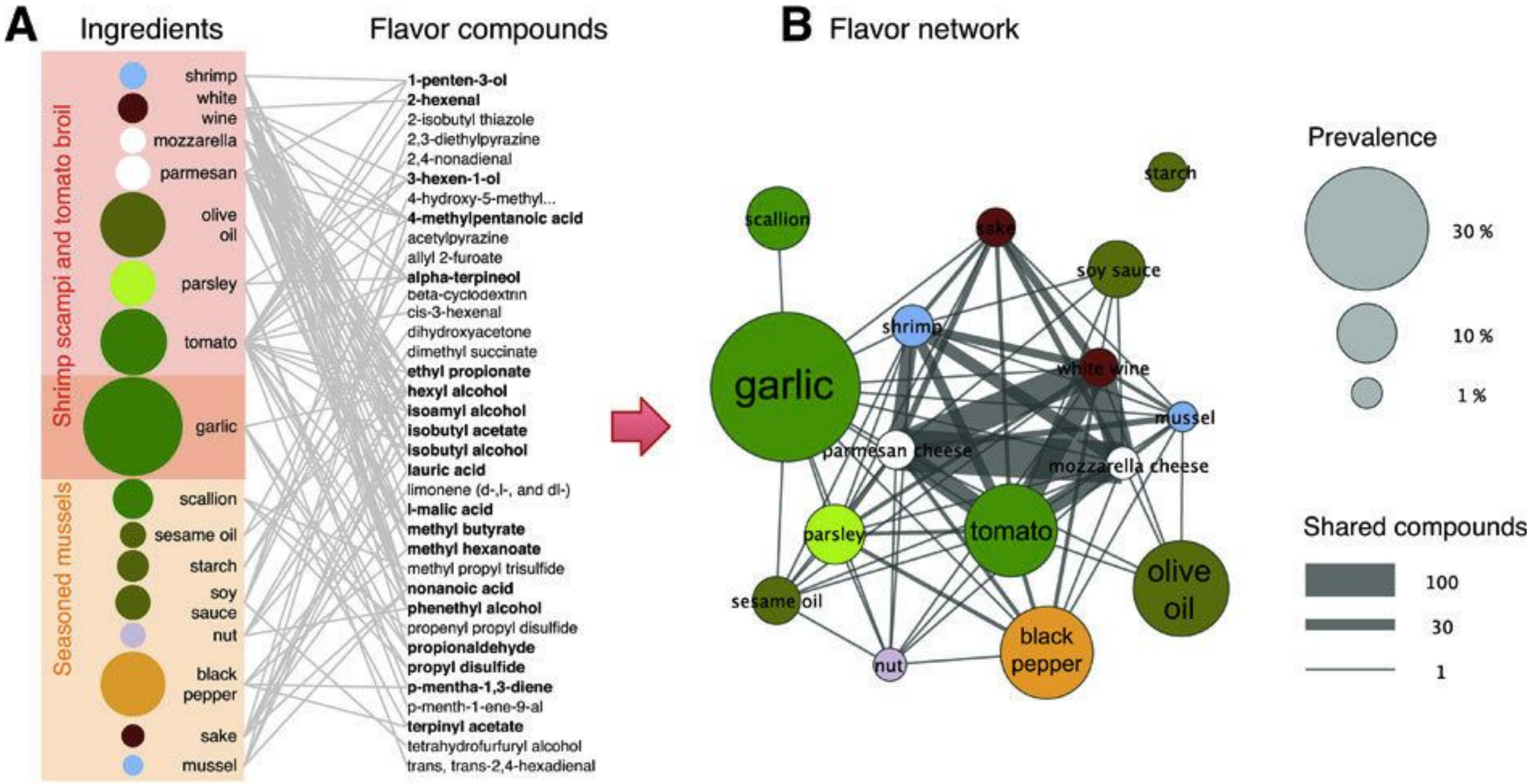




Disease network
(projection)

Gene - DiseNetwork ase Network

Goh, Cusick, Valle, Childs, Vidal & Barabási, PNAS (2007)



Ingredient-Flavor Bipartite Network

Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, A.-L. Barabási
Flavor network and the principles of food pairing,
Scientific Reports 196, (2011).

Summarizing...

Central quantities in Network Science

Degree Distribution

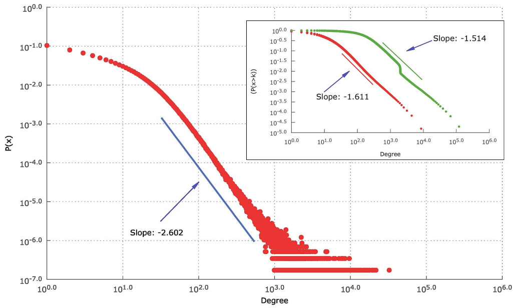
$$P(k)$$

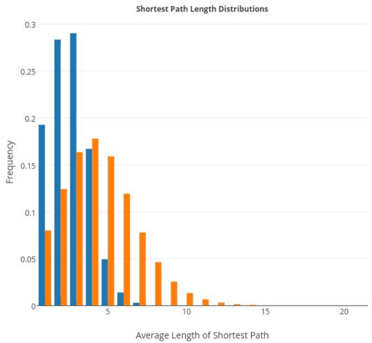
Path length

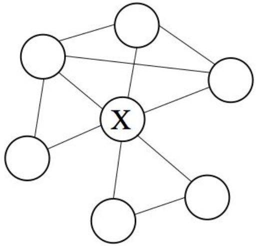
$$\langle d \rangle$$

Clustering Coefficient

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



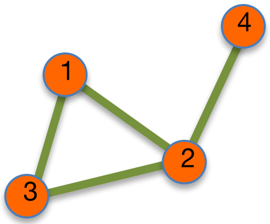




Type of graphs

Directedness

Undirected graph



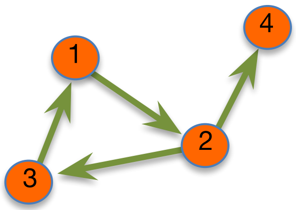
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed graph



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

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

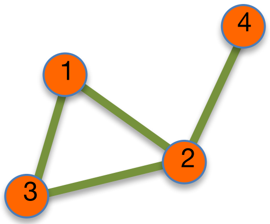
$$L = \sum_{i,j=1}^{N} A_{ij} \qquad < k > = \frac{L}{N}$$

WWW, citation networks

Type of graphs

Weightedness

Unweighted graph



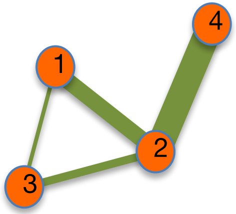
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

protein-protein interactions, WWW

Weighted graph



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

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

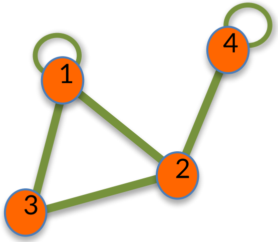
$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k >= \frac{2L}{N}$$

Call Graph, metabolic networks

Type of graphs

Loops & Multigraphs

Self Interactions



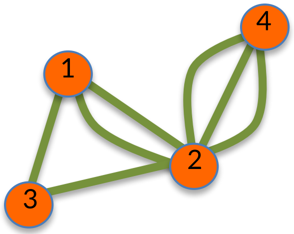
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

protein-protein interactions, WWW

Multigraph (undirected)



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

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

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

Social Network, Collaboration Network

Network	Directed	Weighted	Multigraph	Self-loops
WWW	yes	no	yes	yes
Protein interactions	no	no	no	yes
Collaboration network	no	yes	yes	no
Mobile phone calls	yes	yes	no	no
Facebook Friendship	no	no	no	no

Real Networks can have multiple characteristics

Case Study:

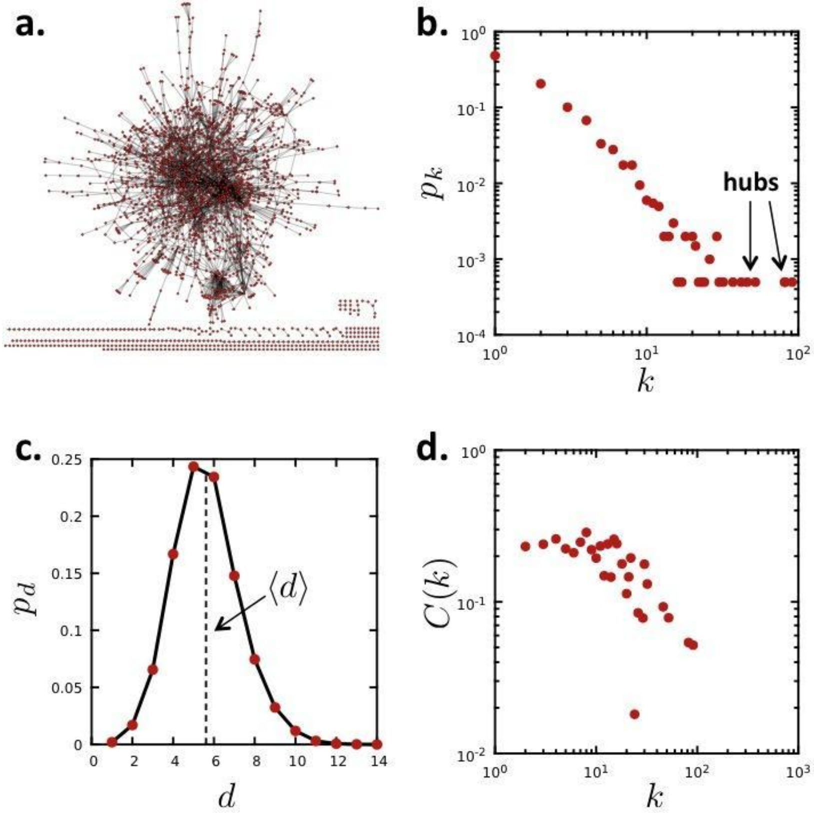
Protein-Protein Interaction Network

Case study

Protein-protein interaction

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

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



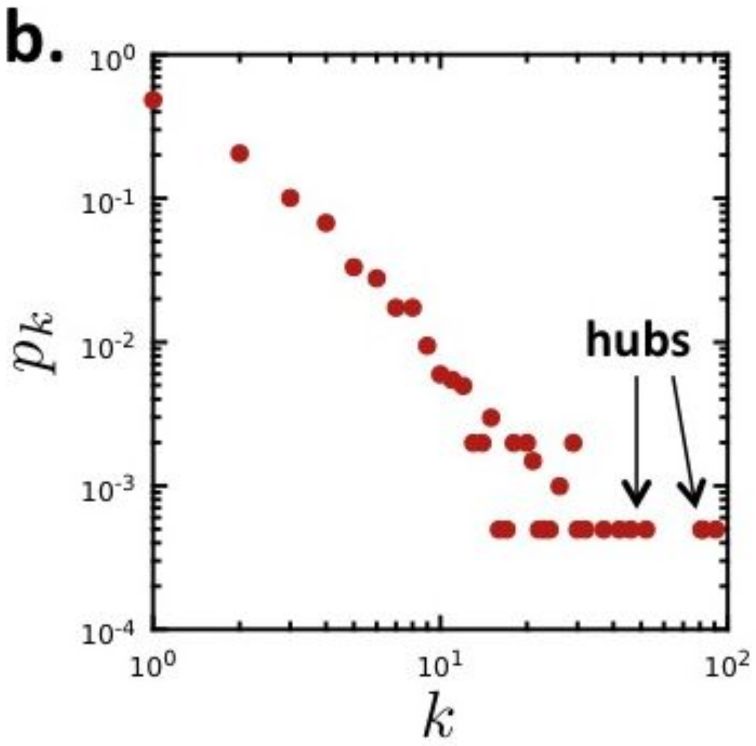
Case study

Protein-protein interaction

p_k is the probability that a node has degree k

N_k = # nodes with degree k

$$p_k = N_k / N$$



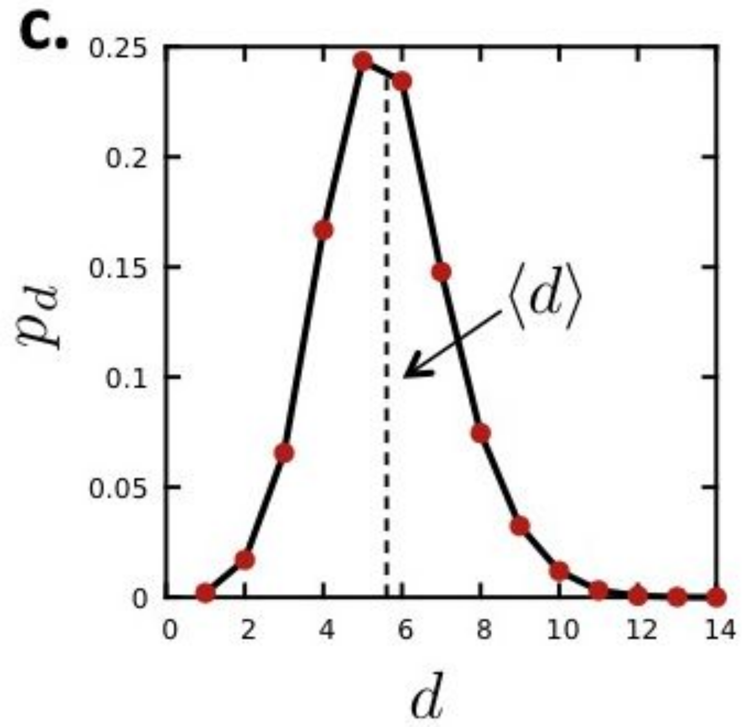
Case study

Protein-protein interaction

Path length distribution

d_{max}=14

<d>=5.61



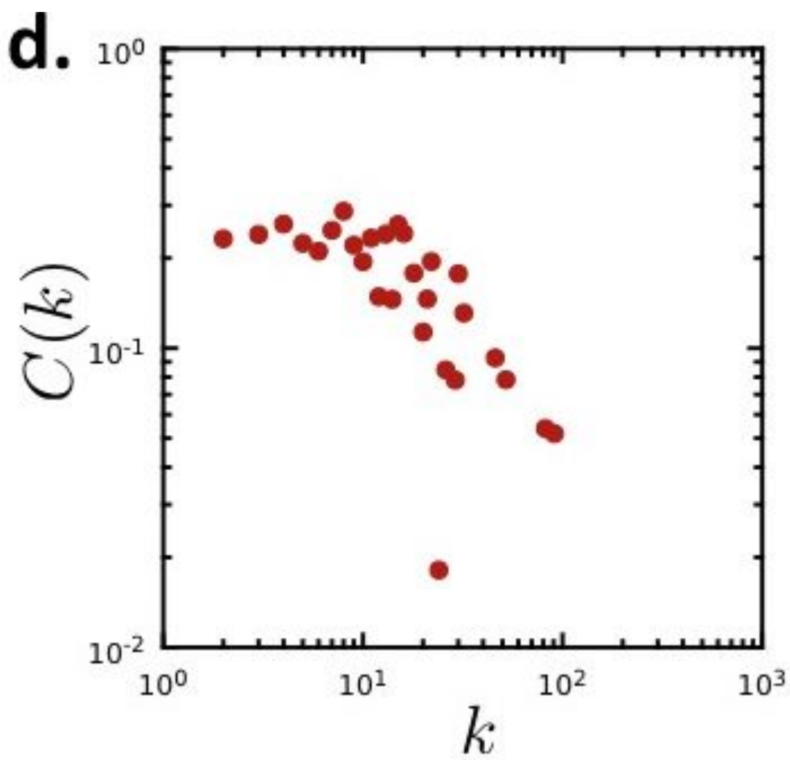
Case study

Protein-protein interaction

Clustering coefficient vs. node degree

Average Clustering Coeff.

<C>=0.12



Chapter 2

Conclusion

Take Away Messages

1. Semantic shapes graph topology
2. Network properties can be measured
3. Degree distribution
4. Paths & Connectivity
5. Clustering Coefficient

Suggested Readings

- Chapter 2 of Barabasi's book
- Chapter 2 of Kleinberg’s book

What’s Next

Chapter 3:
Random Networks

Notebook

Chapter 2: Basic Measures
https://github.com/sna-unipi/SNA_lectures_notebooks