P.h.D. course
Universitat Autonoma de Barcelona, 2022

(**S**ocial) **N**etwork **A**nalysis

Giulio Rossetti

Knowledge Discovery and Data Mining Laboratory (KDD) @ ISTI-CNR

giulio.rossetti@isti.cnr.it

@GiulioRossetti

ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

About me...

Current Position

- Permanent Researcher @ CNR-ISTI, Italy
- External Prof. of “Social Network Analysis” @ University of Pisa
- Coordinator of SNA research @ KDD lab

Research interests

- Complex Networks, Epidemic Modelling, Polluted Information Environments

EU Projects Experience

- WP Leader: SoBigData++
- Unit-PI: HumMingBird, European Patent Office
- Team member: DATASIM, SEEK, CIMPLEX, Track&Know, SAI

Software Libraries (Maintainer)

- NDlib: Network Diffusion
- CDlib: Community Discovery
- DyNetx: Dynamic Network modeling

Giulio Rossetti
CNR-ISTI, Italy
giulio.rossetti@isti.cnr.it

(Tentative) Agenda

Lecture 1: Introduction to Network Analysis: Measures and fundamentals

Lecture 2: Characterizing by contraposition: Real Networks and Synthetic Models

Lecture 3: Micro, Meso & Macro: Different perspectives

Lecture 4: Dynamics of Networks: Topology perturbations

Lecture 5: Dynamics on Networks: Diffusive phenomena

Appendix: Hand-on (Gephi & Python)

Course Materials

Reference textbooks:

- D. Easley, J. Kleinberg:
  Networks, Crowds, and Markets.
- A. L. Barabasi:
  Network Science
- D. Zinoviev:
  Complex Network Analysis in Python
- M. Coscia:
  The Atlas for Aspiring Network Scientists

GitHub Repository:

- Python Tutorials
- Slides
- Links to software resources

shorturl.at/fyAJM

Lecture 1

Introduction to Network Analysis: Measures and fundamentals

Chapter 0

Why should we care about Complex Networks?

Summary

- Complexity
- Real world networks
- Emergence of Network Science

Reading

- Chapter 1 & 2 of Kleinberg's book
- Chapter 1 of Barabasi's book.
- Complexity Explained

Complex

[adj., v. kuh m-pleks, kom-pleks; n. kom-pleks]
adjective

1. Composed of many interconnected parts; compound; composite: a complex highway system.

2. Characterized by a very complicated or involved arrangement of parts, units, etc.: complex machinery.

3. So complicated or intricate as to be hard to understand or deal with: a complex problem.

Source: Dictionary.com

Complexity, a scientific theory which asserts that some systems display behavioral phenomena that are completely inexplicable by any conventional analysis of the systems' constituent parts. These phenomena, commonly referred to as emergent behaviour, seem to occur in many complex systems involving living organisms, such as a stock market or the human brain.

Source: John L. Casti, Encyclopædia Britannica

Complexity

Behind each complex system
there is a network,
that defines the interactions
between the components.

Suggested Reading
Complexity Explained
https://complexityexplained.github.io/

Examples of

Complex Systems

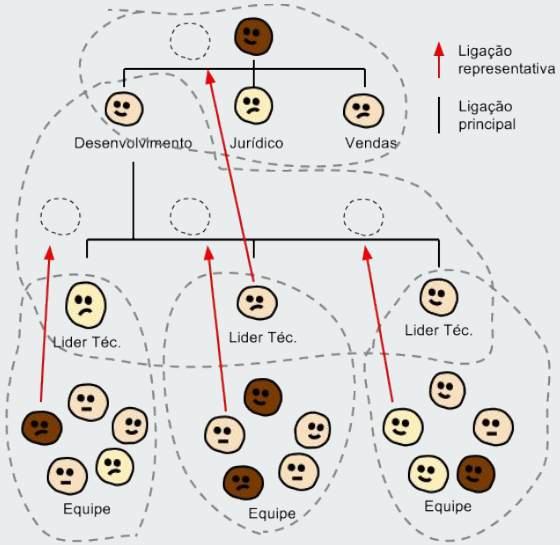
The Facebook “Social Graph”

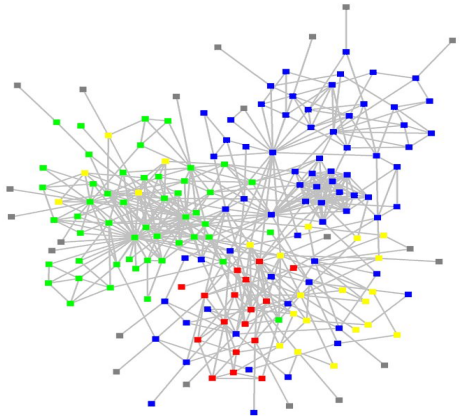
Keith Shepherd's "Sunday Best”.
http://baseballart.com/2010/07/shades-of-greatness-a-story-that-needed-to-be-told/

Examples of

Complex Systems

The structure of an organization





departments

consultants

external experts

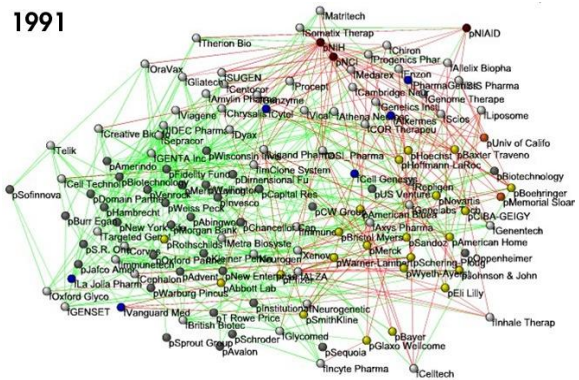
www.orgnet.com

Links:

- Collaborations
- Financial
- R&D

Nodes:

- Companies
- Investment
- Pharma
- Research Labs
- Public
- Biotechnology

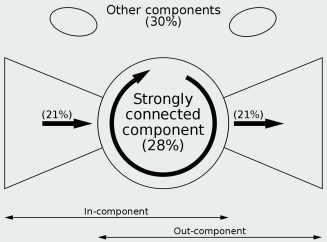


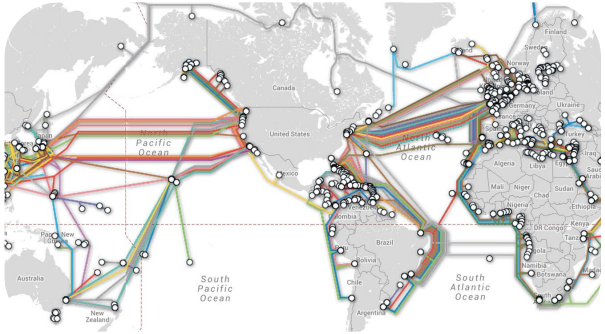
http://ecclectic.ss.uci.edu/~drwhite/Movie

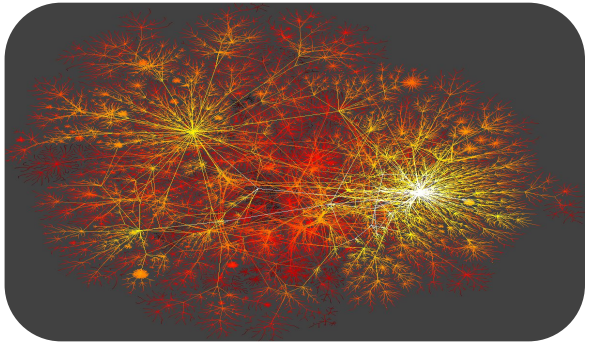
Examples of

Complex Systems

The Internet backbone,
The World Wide Web...







Examples of

Complex Systems

Human Genes

Humans have only about three times as many genes as the fly, so human complexity seems unlikely to come from a sheer quantity of genes.

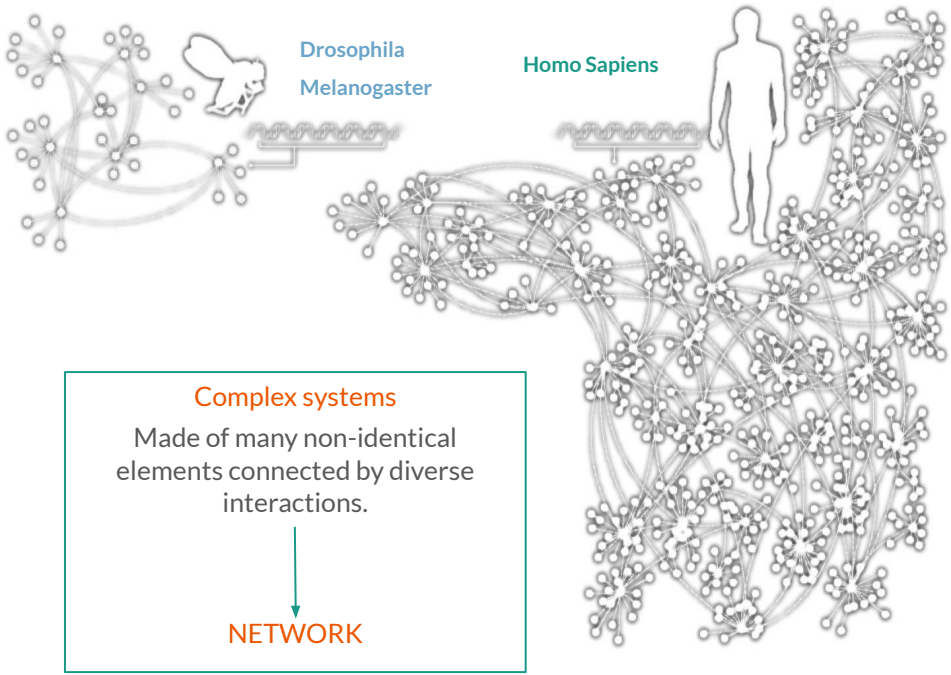
Rather, some scientists suggest, each human has a network with different parts like genes, proteins and groups.

Examples of

Complex Systems

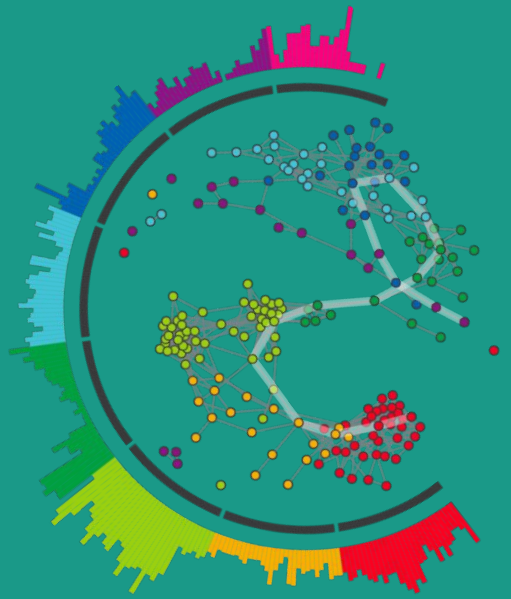
Human Genes (cont'd)

In the generic networks shown, the points represent the elements of each organism's genetic network, and the dotted lines show the interactions between them.



The role of networks

Behind each system studied in complexity there is an intricate wiring diagram, or a **network**, that defines the interactions between the component.



We will never understand **complex system** unless we map out and understand the networks behind them.

Examples of

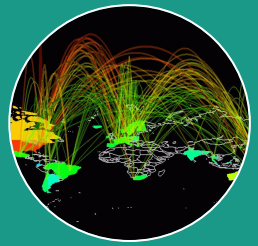
Real world Networks

Type: Social
Nodes: Individuals
Links: Social relationship

NetSci
network science society

Type: Scientific Collaborations
Nodes: Researchers
Links: Co-Authorships

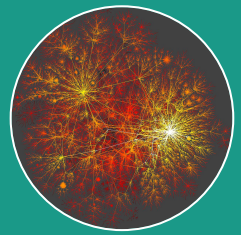
Type: Actor connectivity
Nodes: Actors
Links: Cast jointly



Type: Communication
Nodes: Phones, Airports..
Links: Phone calls, Flights..

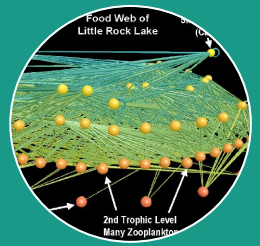
Examples of

Real world Networks (cont'd)



Type: Technological
Nodes: PC, Routers
Links: Physical lines

Type: Scientific Citation
Nodes: Papers
Links: Citations



Type: Biological
Nodes: Species
Links: Trophic interactions

Type: Mobility
Nodes: Individuals, Cars...
Links: Co-Location...

The Emergence of Network Science

The (urgent) need to understand complexity

Despite the challenges complex systems offer us, we cannot afford to not address their behavior, a view increasingly shared both by scientists and policy makers.

Networks are not only essential for this journey, but during the past decade some of the most important advances towards understanding complexity were provided in context of network theory.

Data Availability

1990 C. elegans neural wiring diagram
1998 - Movie Actor Network
1998 - Citation Networks
1999 -World Wide Web
2000 - Metabolic Networks
2001 - PPI network
2008 - OSNs

Universality

The architecture of networks emerging in various domains of science, nature, and technology are more similar to each other than one would have expected.

The Tools of

Modern Network Theory

1. Graph Theory
2. Social Network Theory
3. Statistical Physics
4. Computer Science
5. Biology
6. Statistics

Chapter 1

Networks & Graphs: Basic Measures

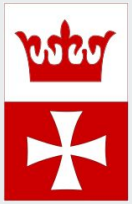
Summary

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

Reading

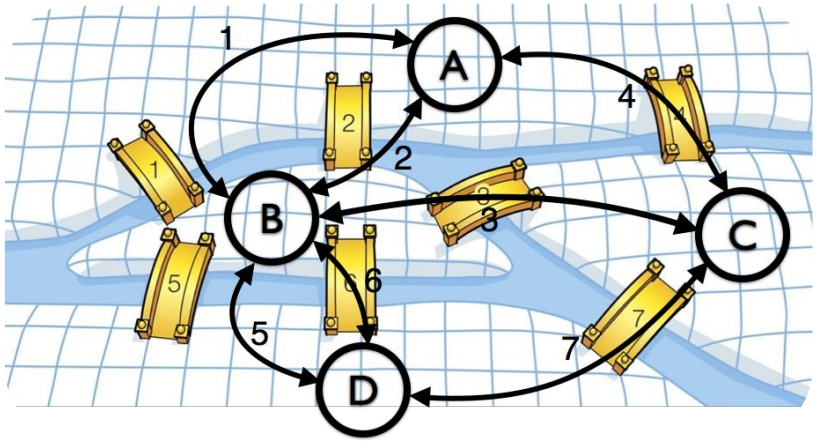
- Chapter 2 of Barabasi's book

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)

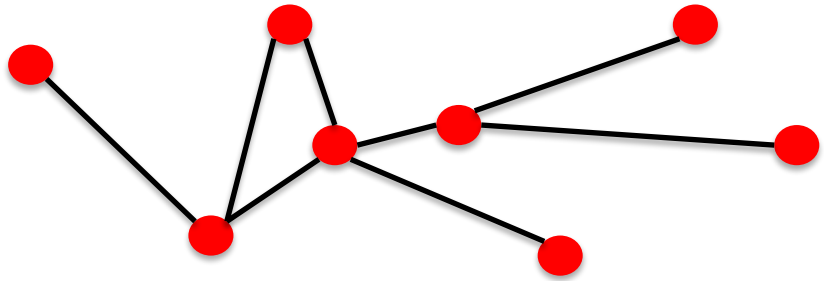
- If a graph has more than two nodes of odd degree, there is no path/cycle that crosses each bridge exactly once.
- If a graph is connected and has no odd degree nodes, it has at least one path.

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)



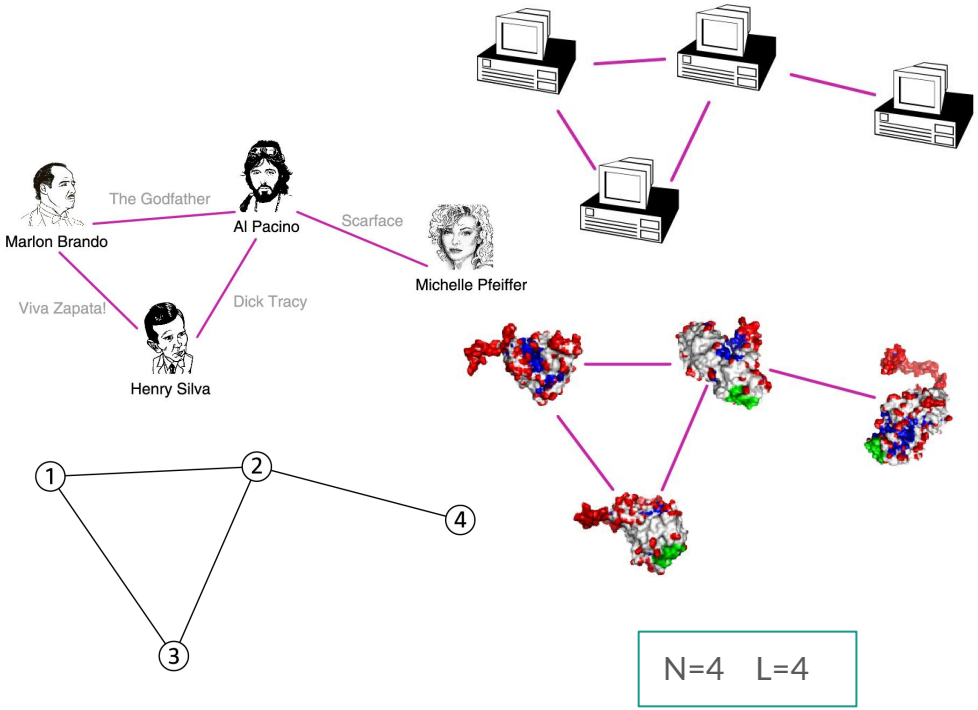
		Symbol
Components	nodes, vertices	N
Interactions	edges, links	L
System	network, graph	(N,L)

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.

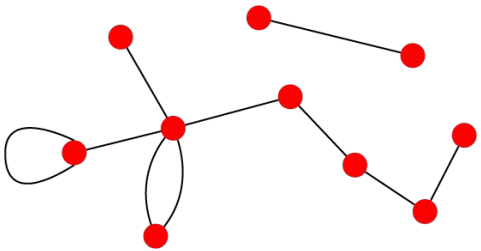
The way we assign the links between a group of individuals will determine the nature of the question we can study.



Directedness

Undirected graphs

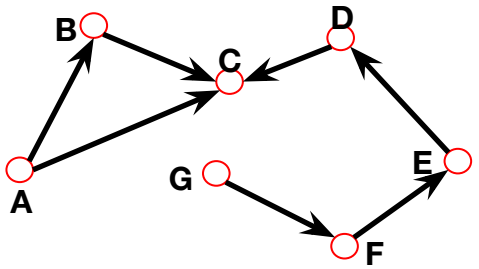
Links: undirected (*symmetrical*)



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

Directed graphs (DiGraphs)

Links: directed (*arcs*).

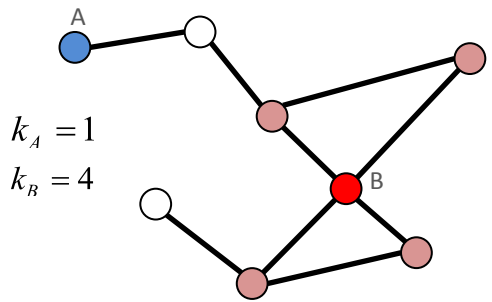


Example of Directed links
- URLs on the www
- Phone calls
- Metabolic reactions

Node Degree

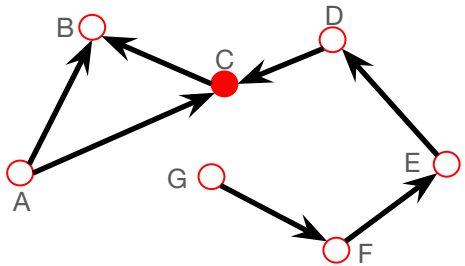
Undirected graphs

the number of links connected to the node



Directed graphs (DiGraphs)

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



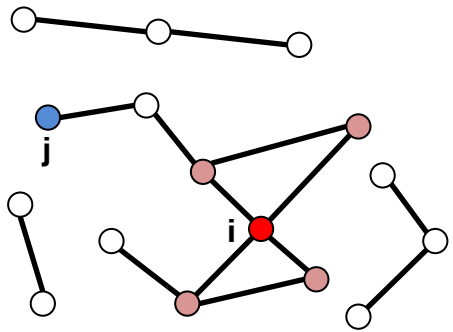
Source: a node with $k^{in} = 0$;

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

Average Degree

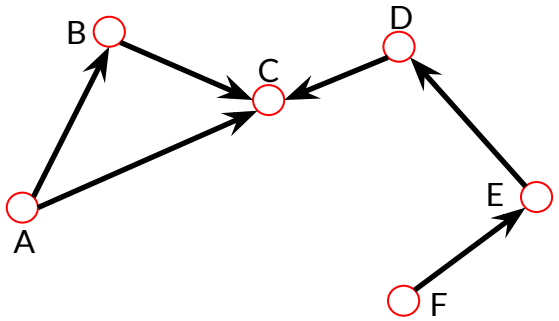
Undirected graphs

N - the number of nodes in the graph



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

Directed graphs (DiGraphs)



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

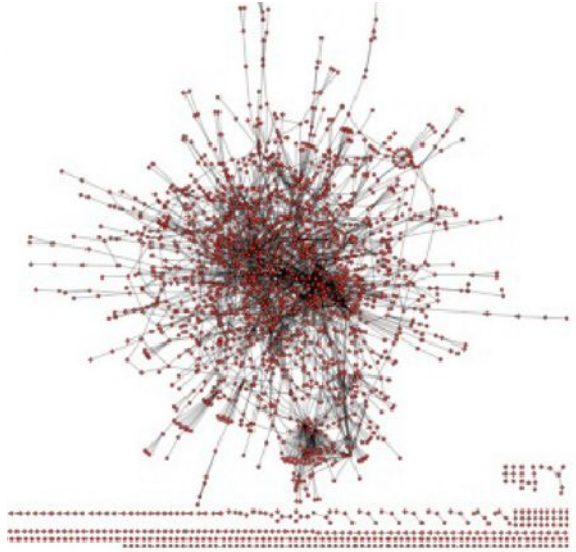
$$\langle \boldsymbol{k}^{in} \rangle = \langle \boldsymbol{k}^{out} \rangle \qquad \langle k \rangle = \frac{L}{N}$$

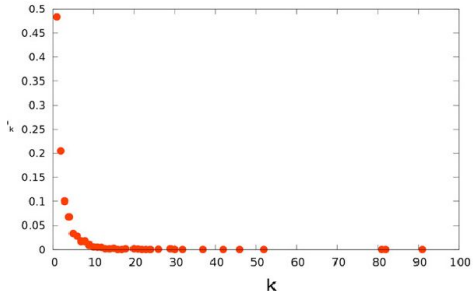
Degree Distribution

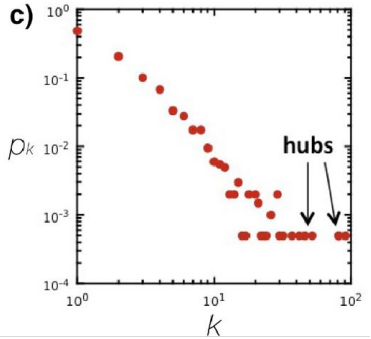
P(k): probability that a randomly chosen node has degree *k*

N_k = # nodes with degree k

$P(k) = N_k / N$ ➔ plot

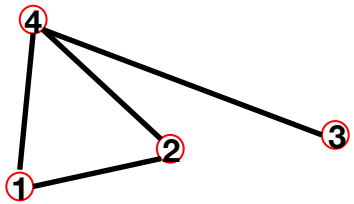






Adjacency matrix

Undirected graphs



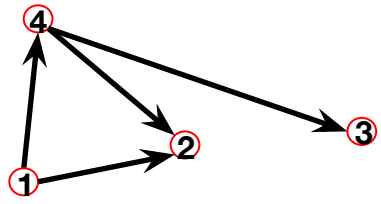
$$A_{ij} = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

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

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

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

Directed graphs (DiGraphs)



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

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

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

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

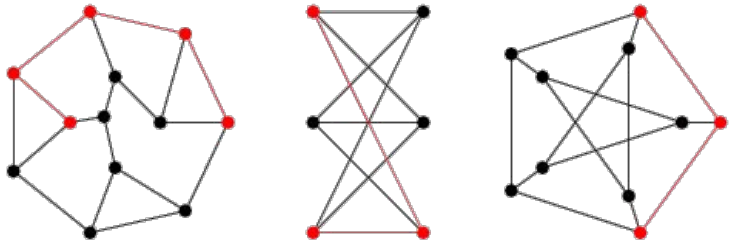
Paths

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

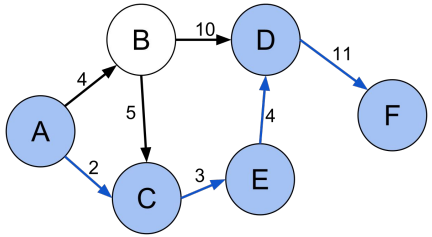
$P_{i0,in}$ of length *n* between nodes i_0 and i_n is an ordered collection of *n+1* nodes and *n* links

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

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



Examples of phats in an undirected graph.

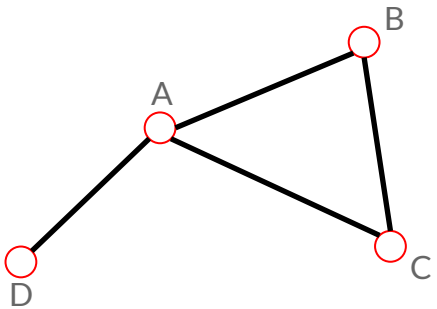


In a directed graph, the path can follow only the direction of an arrow.

Distance in a Graph

Undirected graphs

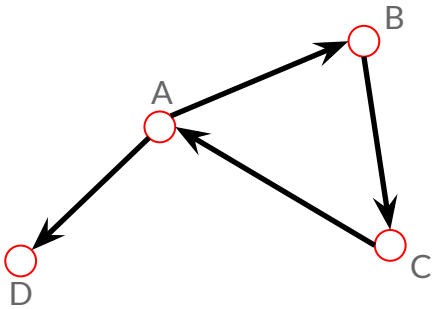
The *distance* (*shortest path*, *geodesic path*) between two nodes is defined as the number of edges along the shortest path connecting them.



*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).

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:

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

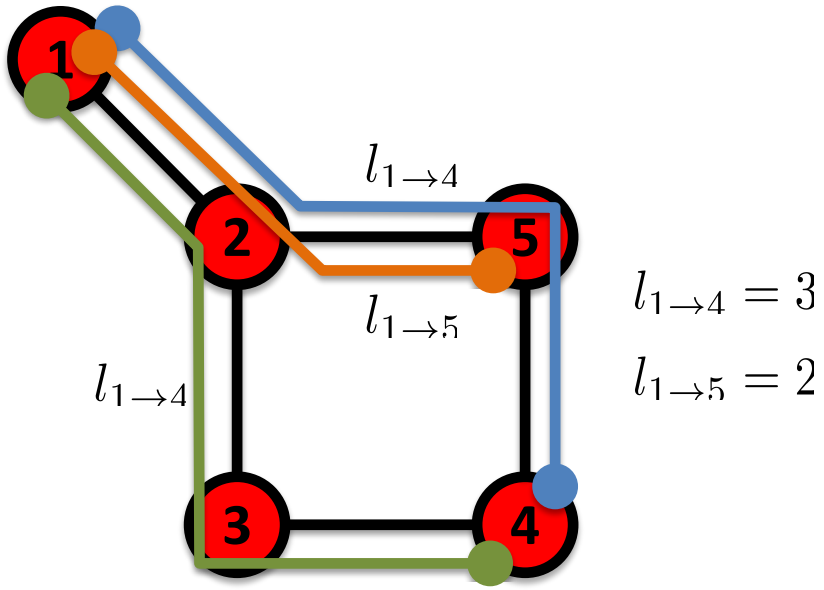
where d_{ij} is the distance from node *i* to node j

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

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

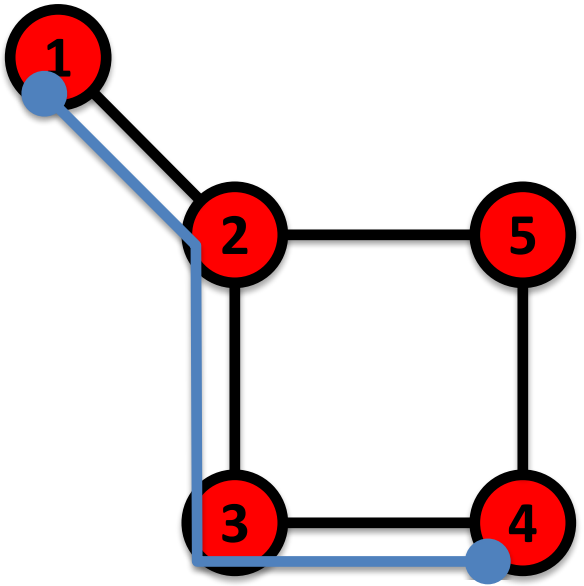
Paths: a summary

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



Paths: a summary

Diameter
The longest shortest path in a graph.

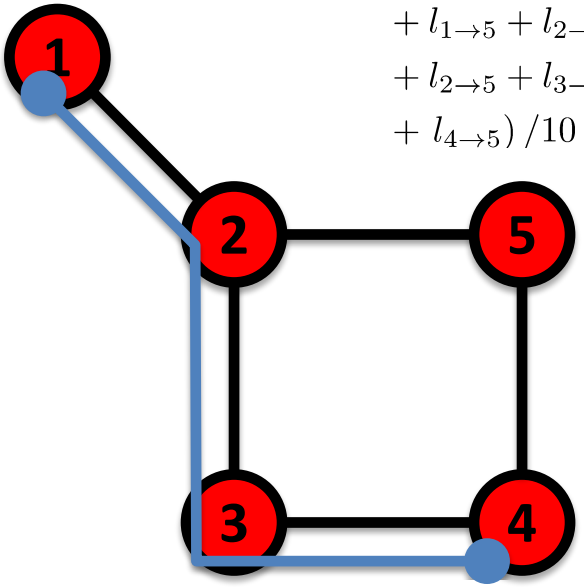


$$l_{1 \to 4} = 3$$

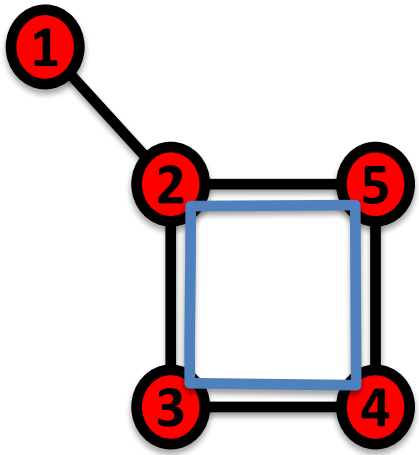
Paths: a summary

Average Path Length
The average of the shortest paths for all pairs of nodes.

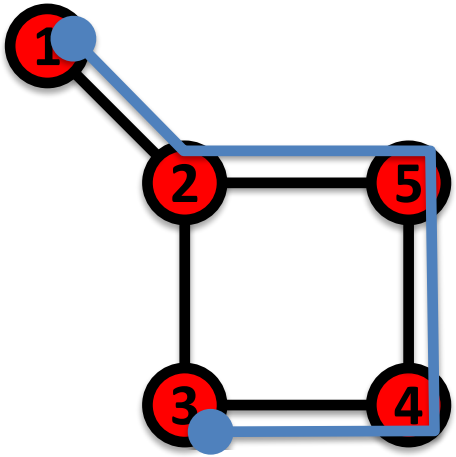
$$(l_{1\to2} + l_{1\to3} + l_{1\to4} + \\ + l_{1\to5} + l_{2\to3} + l_{2\to4} + \\ + l_{2\to5} + l_{3\to4} + l_{3\to5} + \\ + l_{4\to5})\,/10 = 1.6$$



Paths: a summary



Cycle
A path with the same start and end node.

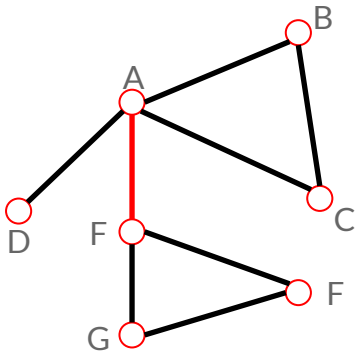


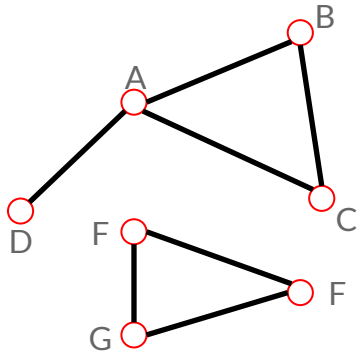
Self-Avoiding Path
A path that does not intersect itself.

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.



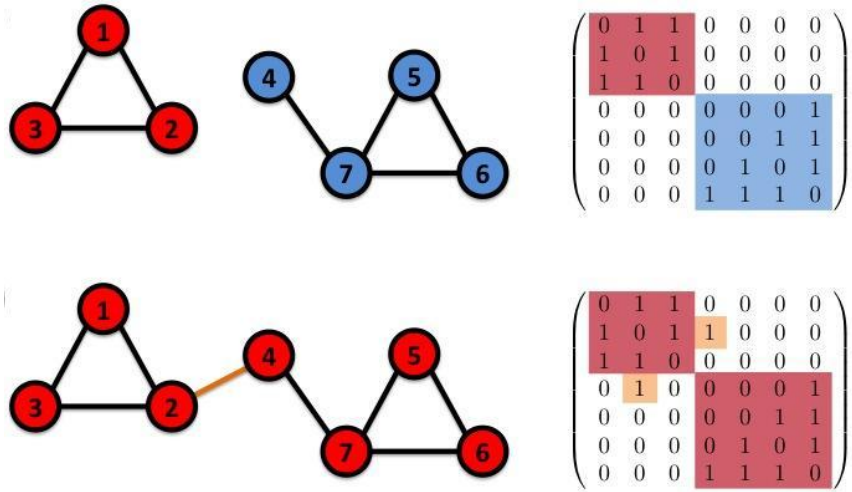


Bridge:
if we erase it, the graph becomes disconnected.
Example (A,F)

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.

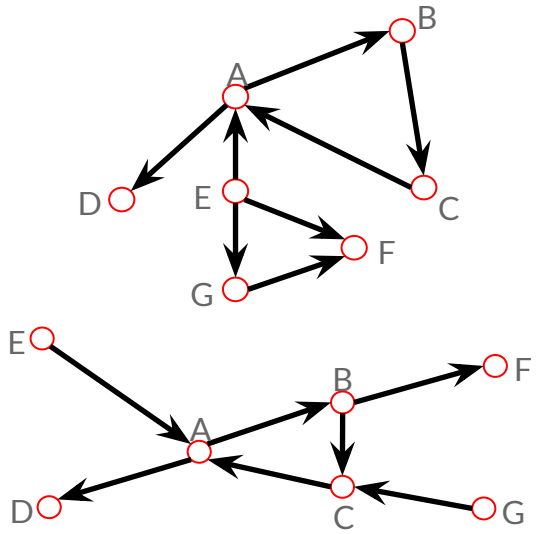


Connectivity of directed graphs

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.

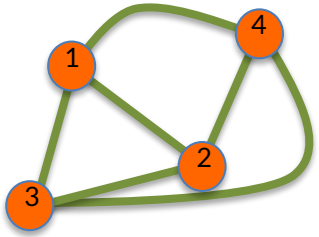
Complete Graph

A graph with degree

$$L=L_{max}$$

is called a complete graph, and its average degree is

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



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

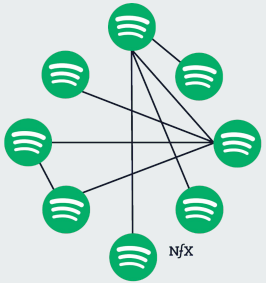
The maximum number of links an undirected network of N nodes can have is:

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

What about directed networks?

Network Density

Ratio of existing edges over possible ones.

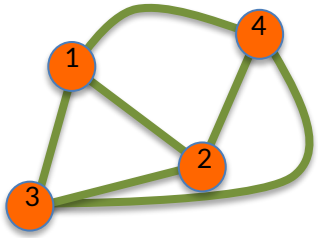


Low Density

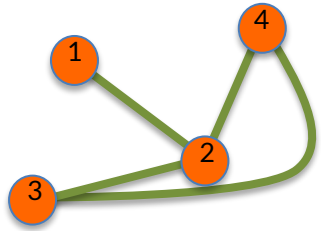
High Density

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

Examples



Complete graph
d(G) = 1



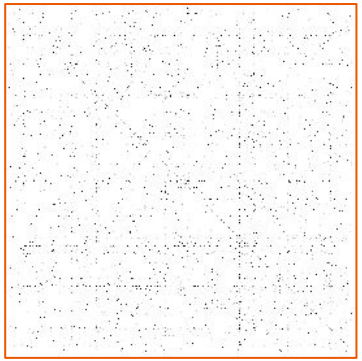
Sparse graph
d(G) = 2/3

Most networks observed in real systems are sparse

$L << L_{max}$

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

$d(G) << 1$



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;	$L=6\ 10^6$	$L_{max}=1.8\ 10^{13}$	$\langle k \rangle=28.78$

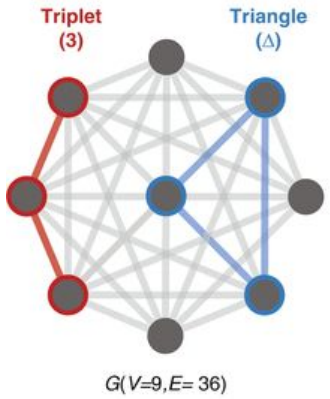
(Source: Albert, Barabasi, RMP2002)

Clustering Coefficient

How “clustered” is my network?

Global Clustering coefficient

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



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

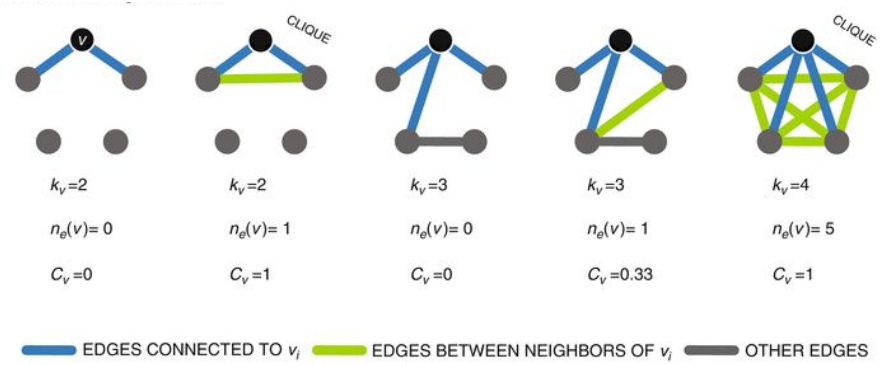
Watts & Strogatz, Nature (1998)

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)

Summarizing...

Central quantities in Network Science

Degree Distribution

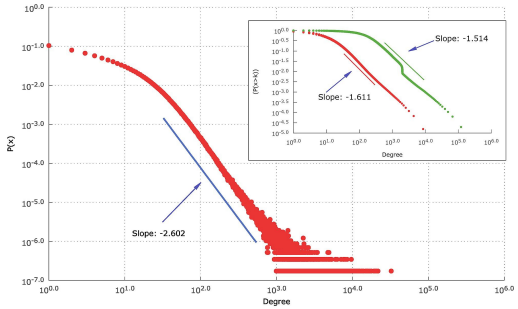
$P(k)$

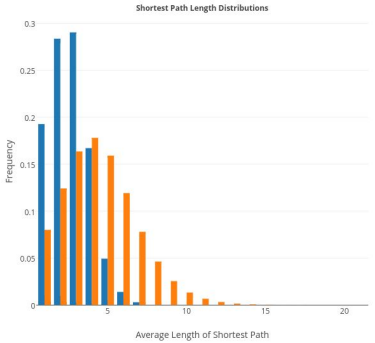
Path length

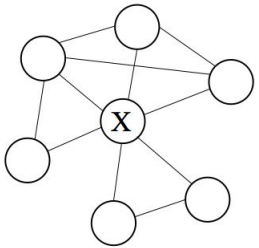
$\langle d \rangle$

Clustering Coefficient

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



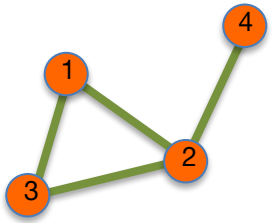




Type of graphs

Directedness

Undirected graph



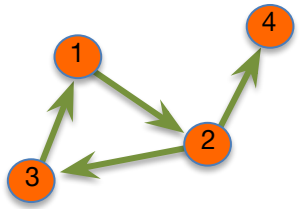
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

Actor network, protein-protein interactions

Directed graph



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

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

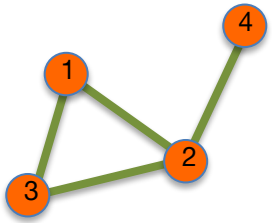
$$L = \sum_{i,j=1}^{N} A_{ij} \qquad < k >= \frac{L}{N}$$

WWW, citation networks

Type of graphs

Weightedness

Unweighted graph



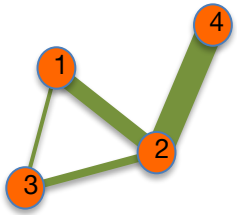
$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

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

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

protein-protein interactions, WWW

Weighted graph



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

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

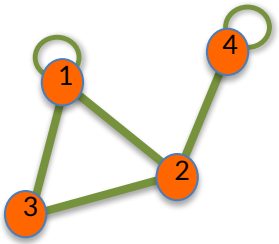
$$L = \frac{1}{2}\sum_{i,j=1}^{N} nonzero(A_{ij}) \qquad < k >= \frac{2L}{N}$$

Call Graph, metabolic networks

Type of graphs

Loops & Multigraphs

Self Interactions



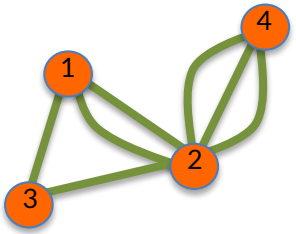
$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

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

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

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

Chapter 1

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

Lecture 2:
Characterizing by contraposition: Real Networks and Synthetic Models