

Introduction to Graph Theory

Proteomes Interactomes and Biological Networks

Emidio Capriotti

<http://biofold.org/>



**Biomolecules
Folding and
Disease**

Department of Pharmacy and
Biotechnology (FaBiT)
University of Bologna



Historical Perspective

With the **Seven Bridges of Königsberg** problem, Euler in 1737 laid the foundations of the graph theory.

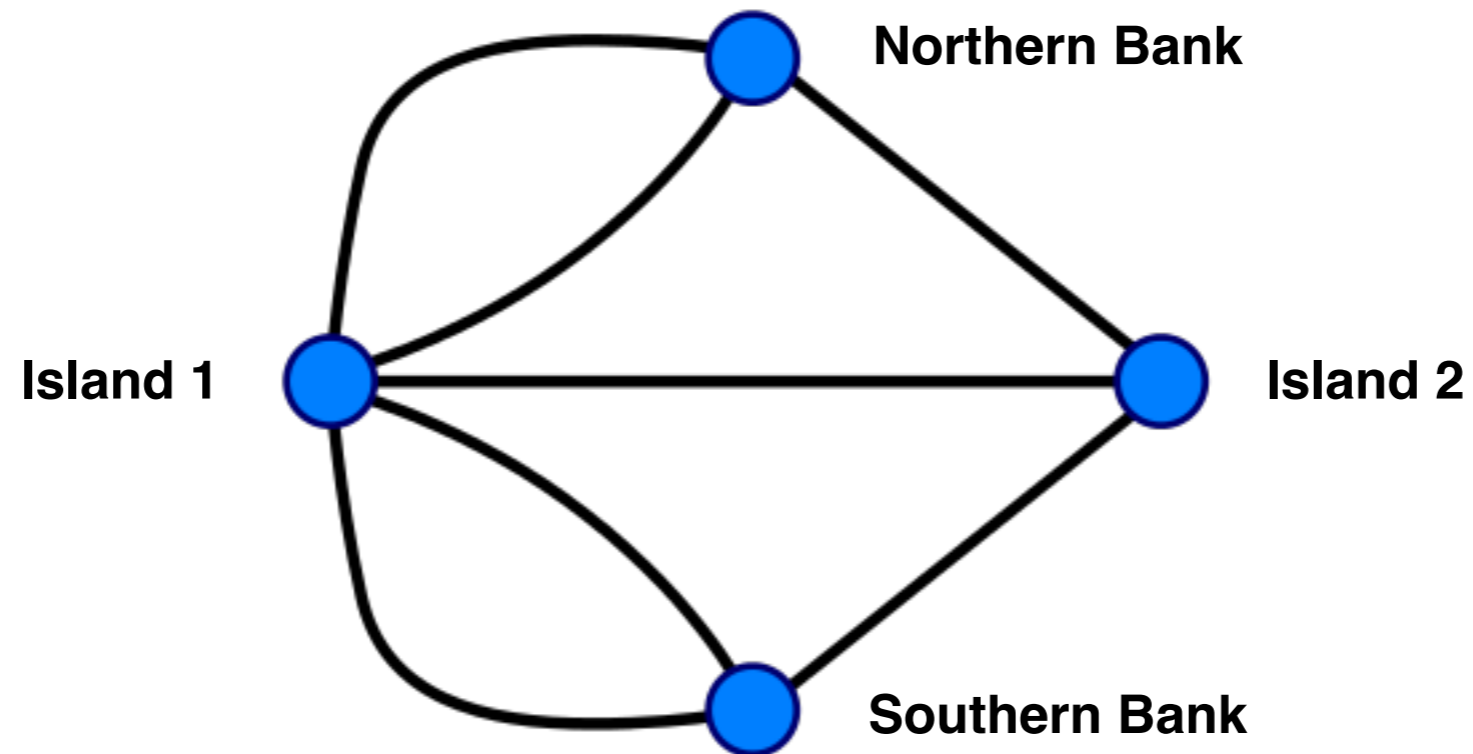


Simon Kneebone – simonkneebone.com

- Find path (Eulerian Path) that **traverses all the Pregel's bridges**.
- Find walk (Eulerian Circuit) that **traverses all the Pregel's bridges** and has **the same starting and ending point**.

Solution

Describe the problem as a graph where the **nodes represent the 4 locations** and the **edges correspond to the bridges**



Eulerian path exists only if **zero or 2 nodes** are connected by an **odd number of bridges**.

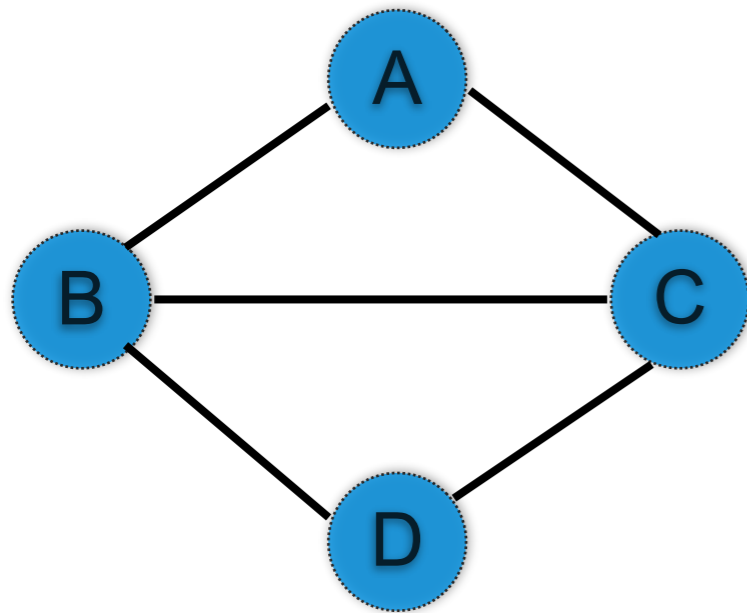
Eulerian circuit exists only if **zero nodes** are connected by an **odd number of bridges**.

Graph Definition

A graph is a pair $G=(V,E)$ consisting of two sets:

- V is a set of elements called **Nodes or Vertices**.
- E is a set of pairs (v_i,v_j) where $v_i \in V$ and $v_j \in V$.

The pairs E are links between two nodes and are called **Edges**

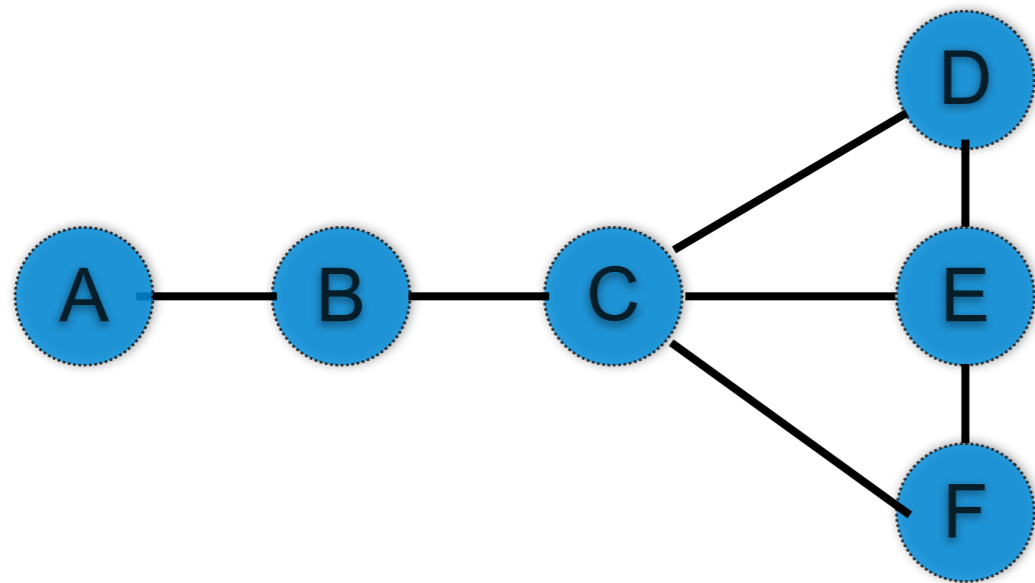


$$V = \{A; B; C; D\}$$

$$E = \{(A,B); (A,C); (B,C); (B,D); (C,D)\}$$

Undirected Graph

Undirected graph is a network where the relationship between nodes are symmetric.

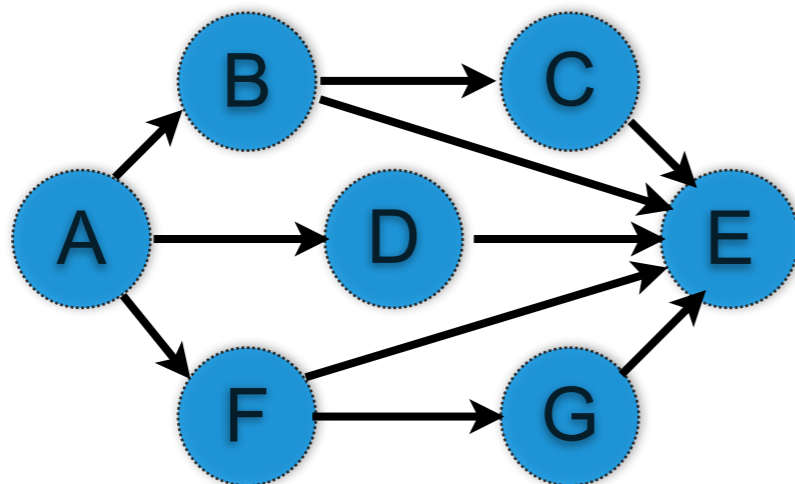
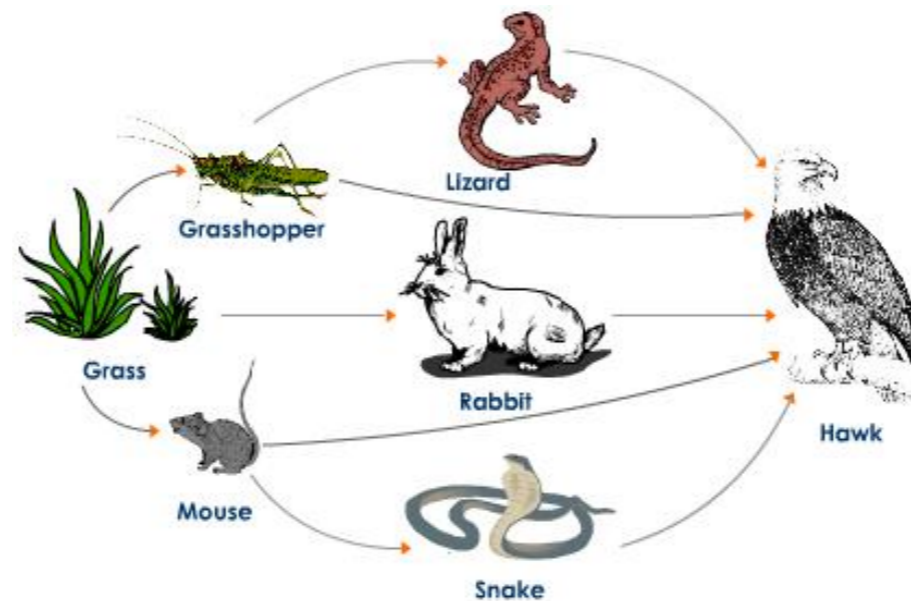


$V = \{\text{Group of People}\}$

$E = \{\text{Pairs of Friends}\}$

Directed Graph

Directed graph is a network where the relationship between nodes are asymmetric. In this case the edges are directed lines.

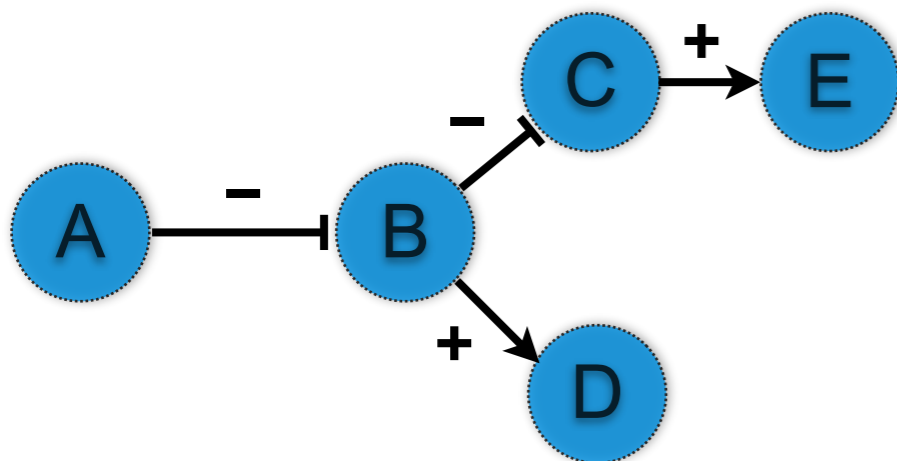
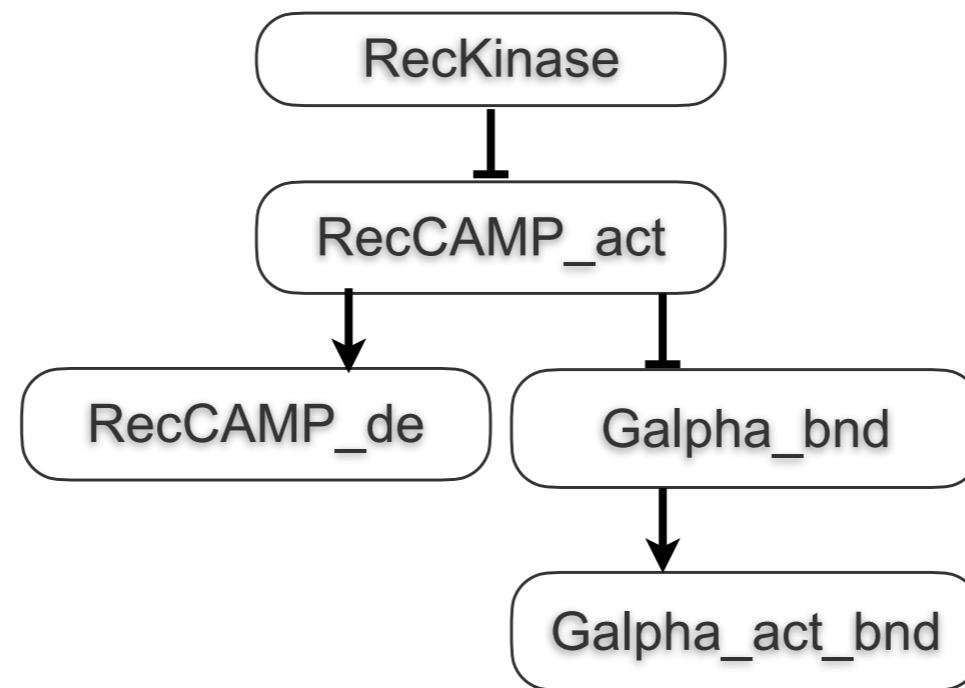


$V = \{\text{Group of Animals}\}$

$E = \{\text{Pray/Predator Relationships}\}$

Signed Directed Graph

Signed Directed graph is a network where the relationship between nodes are asymmetric and have positive or negative associated signs



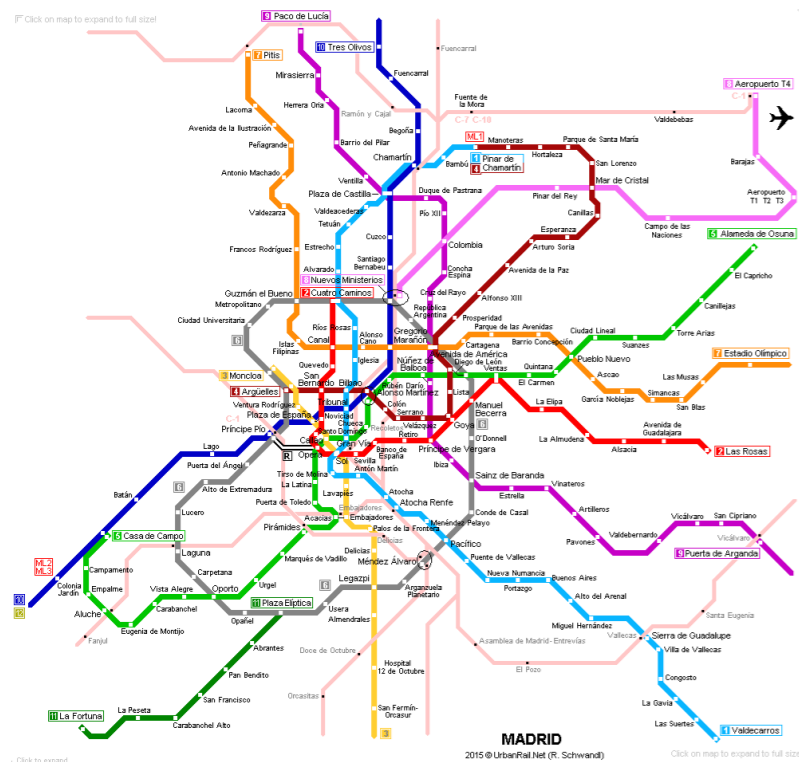
$V = \{\text{Group of Genes}\}$

$E = \{\text{Activation/Inhibition Relationships}\}$

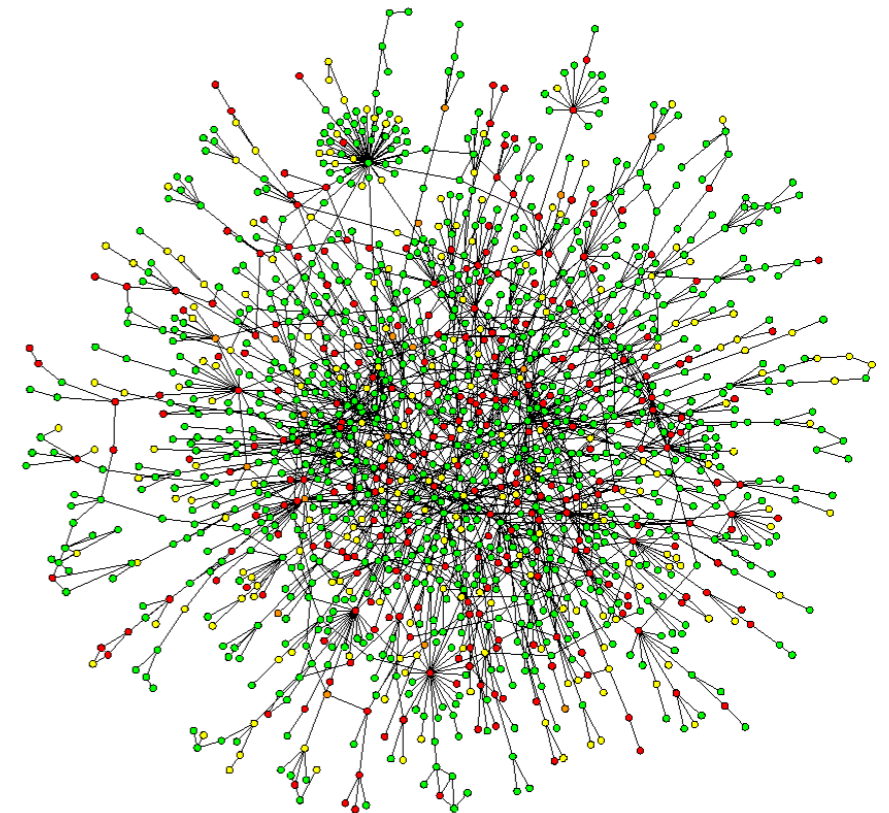
Graph and Networks

Graphs can be used to represent any observed network.

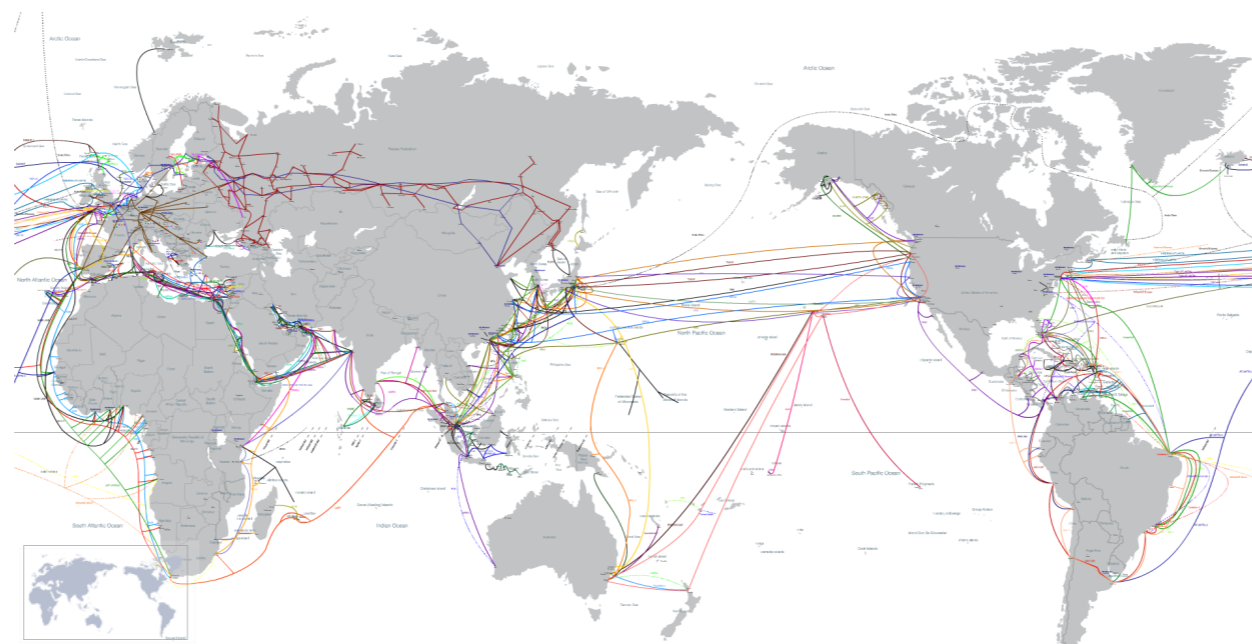
Networks in nature tend to be highly complex



Madrid Metro



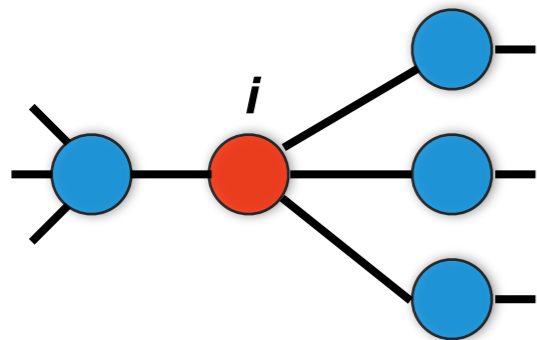
Internet connections



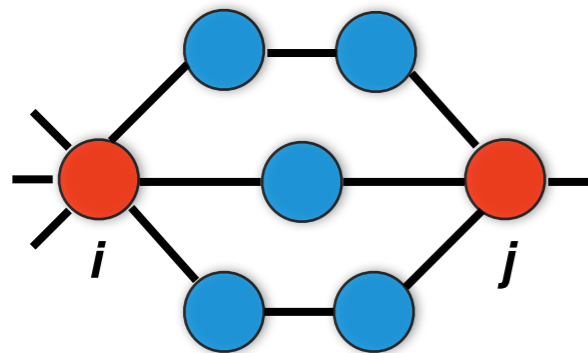
Yeast interactome

Network properties (I)

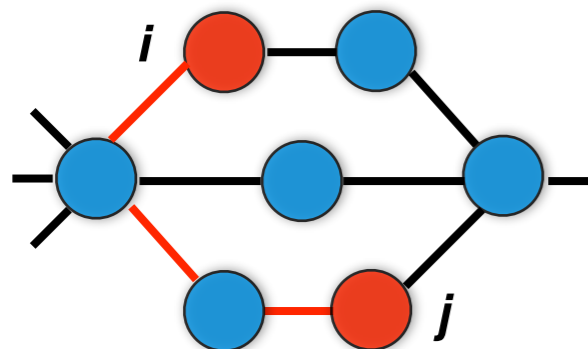
The **topology** of the network **defines its properties**. The level of **connectivity** among the nodes **depends on the number of edges**.



Degree $k_i =$ number of links connected to node i



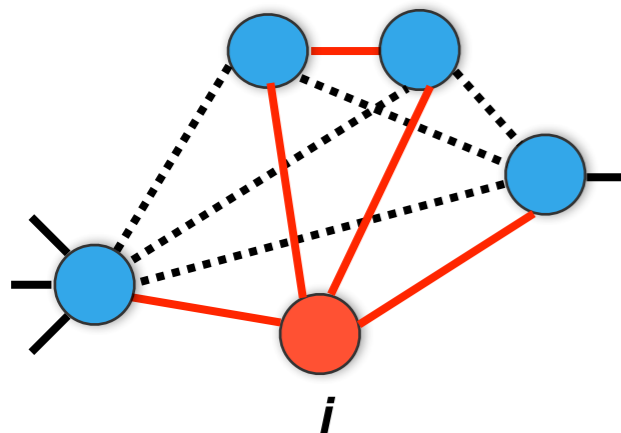
Distance $d_{ij} =$ shortest path between nodes i and j



Diameter $D =$ longest path between all pairs of nodes

Network properties (II)

The **topology** of the network **defines its properties**. The level of **connectivity** among the nodes **depends on the number of edges**.

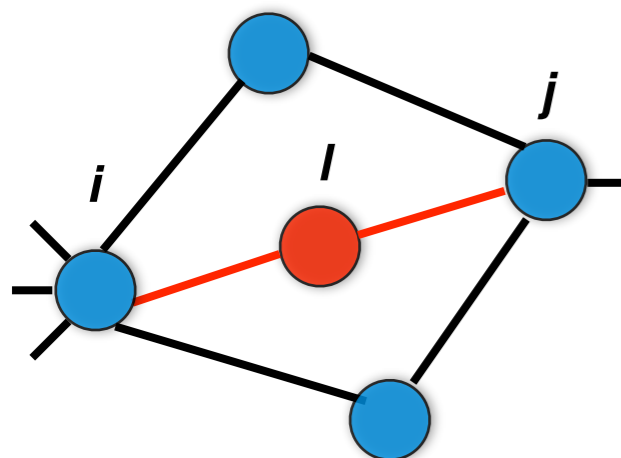


Transitivity
or
Clustering
Coefficient

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

k_i = number of nodes connected to i

e_i = number of edges between the k_i nodes



Betweenness

$$g_l = \sum_{i \neq l \neq j} \frac{\sigma_{ij}(l)}{\sigma_{ij}}$$

σ_{ij} = number of shortest path between i and j

$\sigma_{ij}(l)$ = number of shortest path passing through node l

Types of Network

The **topology** of the network depends on the **distribution of the degree** for all the nodes.

We can define three types of network:

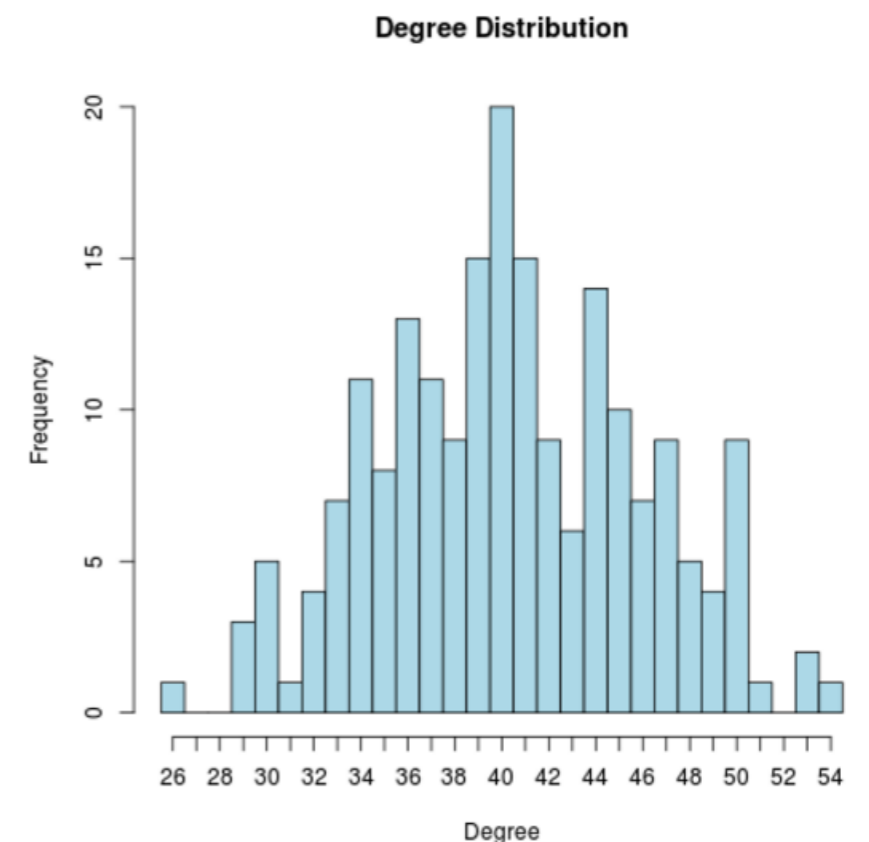
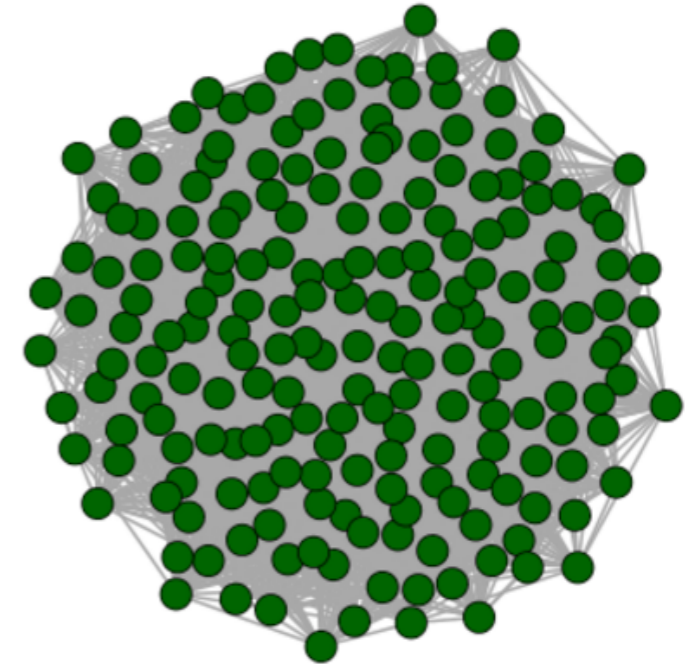
- **Random network**: generated by a constant probability of having an edge between two nodes.
- **Small-world network**: when the degrees follow a Poisson distribution
- **Scale-Free network**: the degrees follow a Power Law distribution

Random Network

Can be generated by Erdős–Rényi model which assume a **constant probability of generating edges** between nodes.

- High node degree \Rightarrow **low average path length**
- Degree **distribution tends to be a Gaussian**
- High Transitivity
- Small Betweenness

Degree = 40.3
Transitivity = 0.2
Betweenness = 79.3



Small-World Network

Generated by a Watts–Strogatz model.

- Low node degree \Rightarrow “Six degrees of separation”
- Degree follows a Poisson distribution
- Low Transitivity than random
- Higher betweenness than random

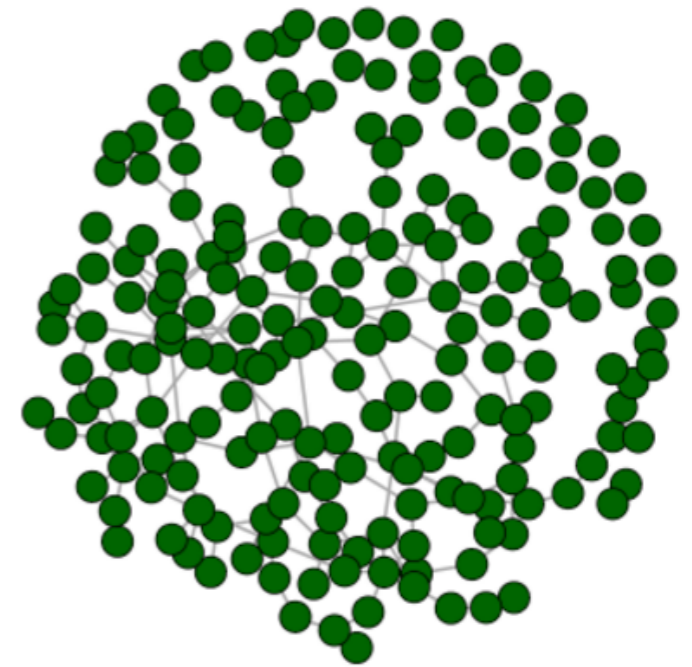
$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

λ = the average value of the distribution
 k = number of observed events

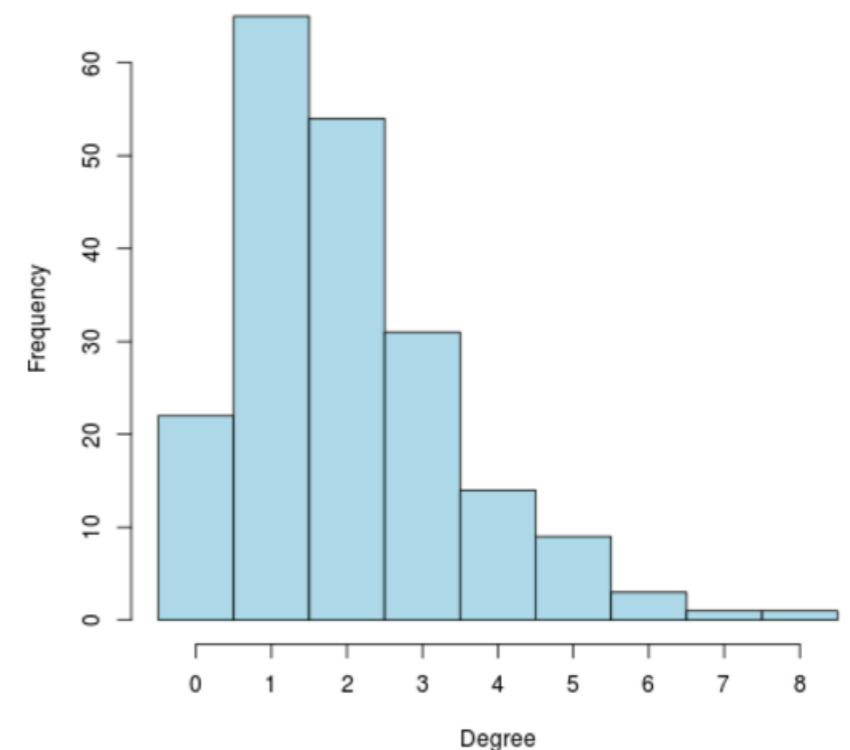
Degree = 2

Transitivity = 0.01

Betweenness = 394.9



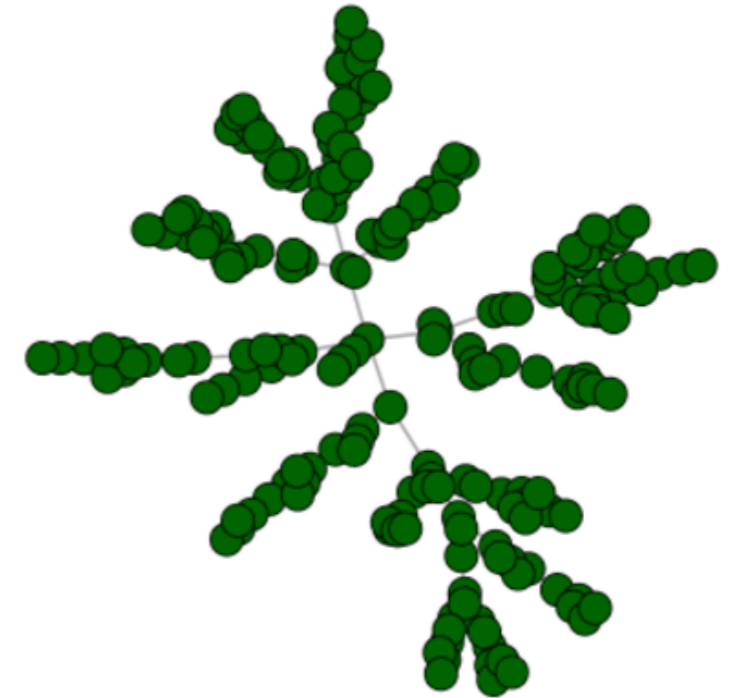
Degree Distribution



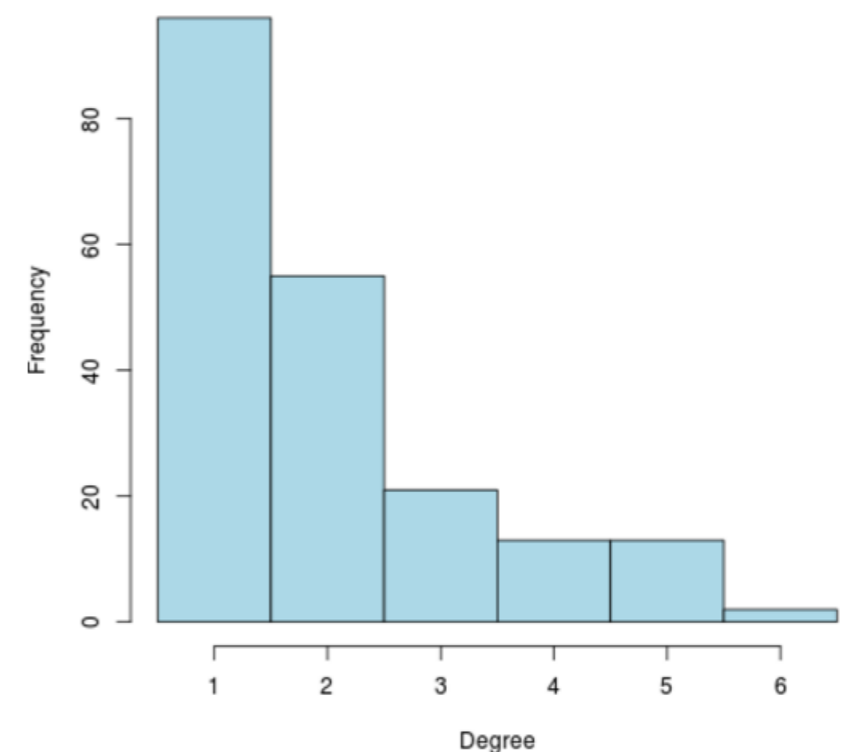
Scale-Free Network

Generated by the Barabasi-Albert model.

- Smallest degree
- Degree follows a Power Law distribution
- Lowest Transitivity
- Highest Betweenness



Degree Distribution



$$p(k) = Ax^{-k}$$

x = is a constant
 k = number of observed events

Degree = 2

Transitivity = 0

Betweenness = 753.4

Biological Network

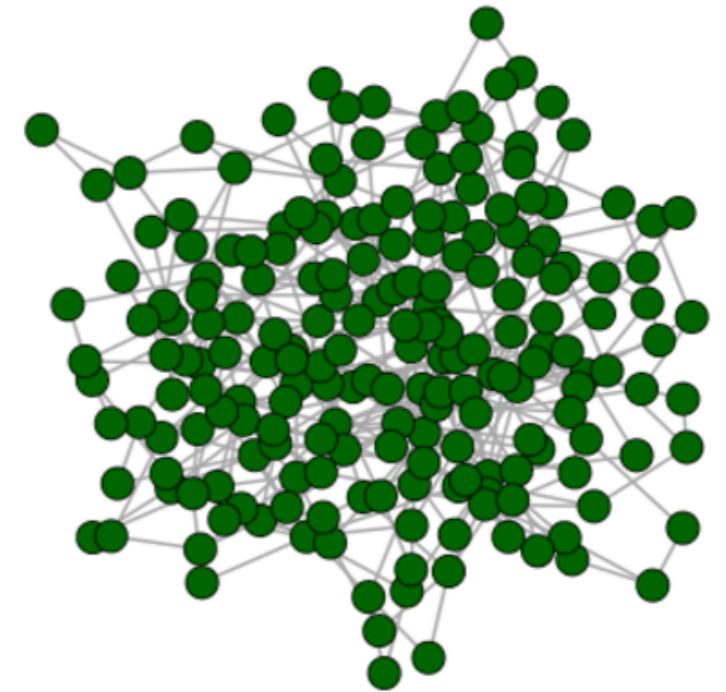
Similar to Small-World and Scale-Free networks

- Small degree
- Average path length proportional to $\ln(\ln(\#nodes))$
- Transitivity high than Small-World and Scale Free
- Betweenness lower than Small-World and Scale Free

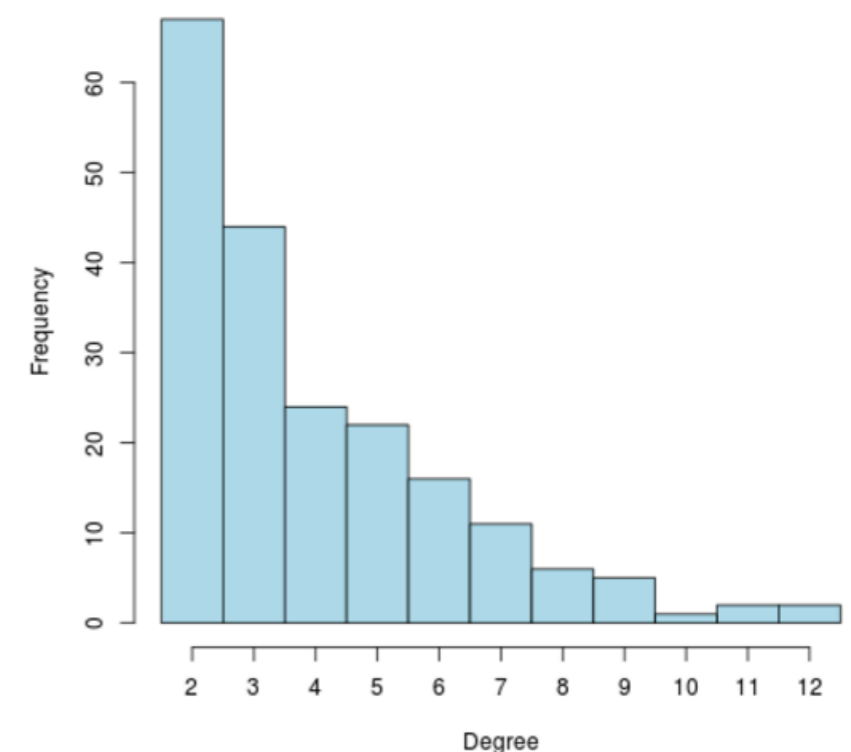
Degree = 4.0

Transitivity = 0.04

Betweenness = 290.4

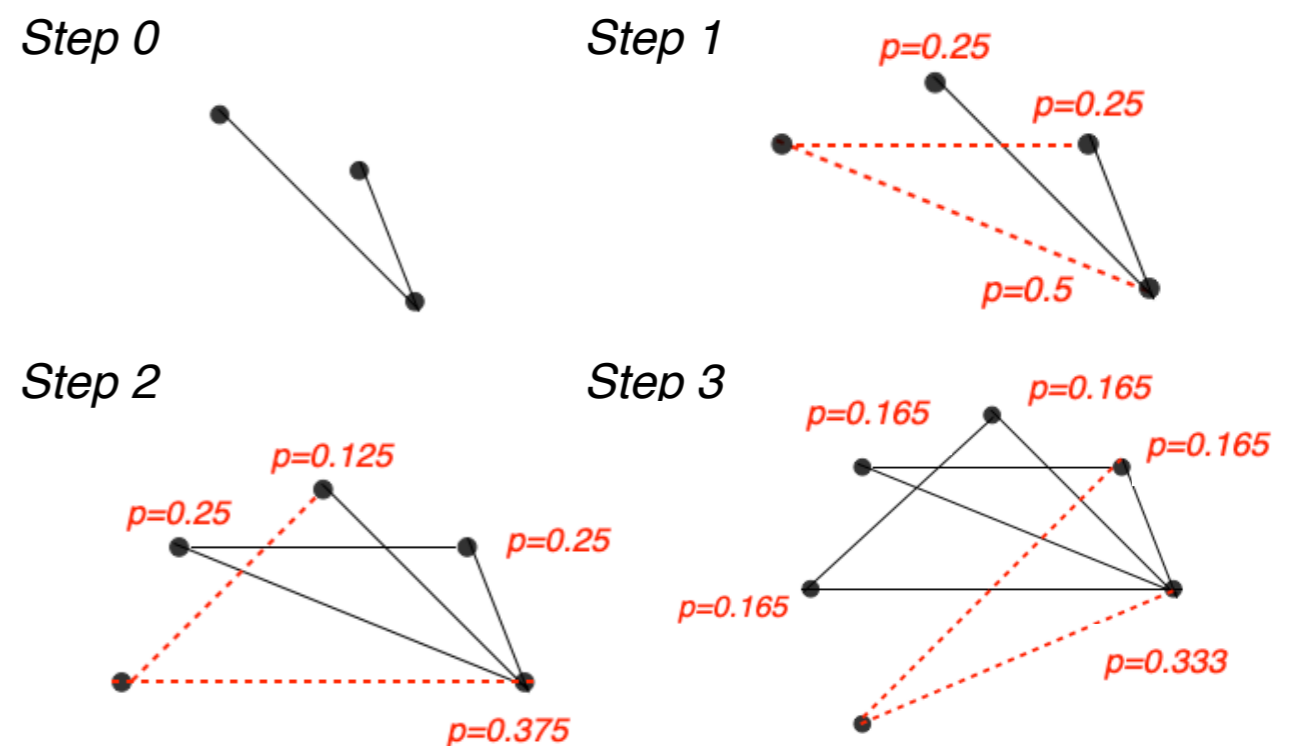


Degree Distribution



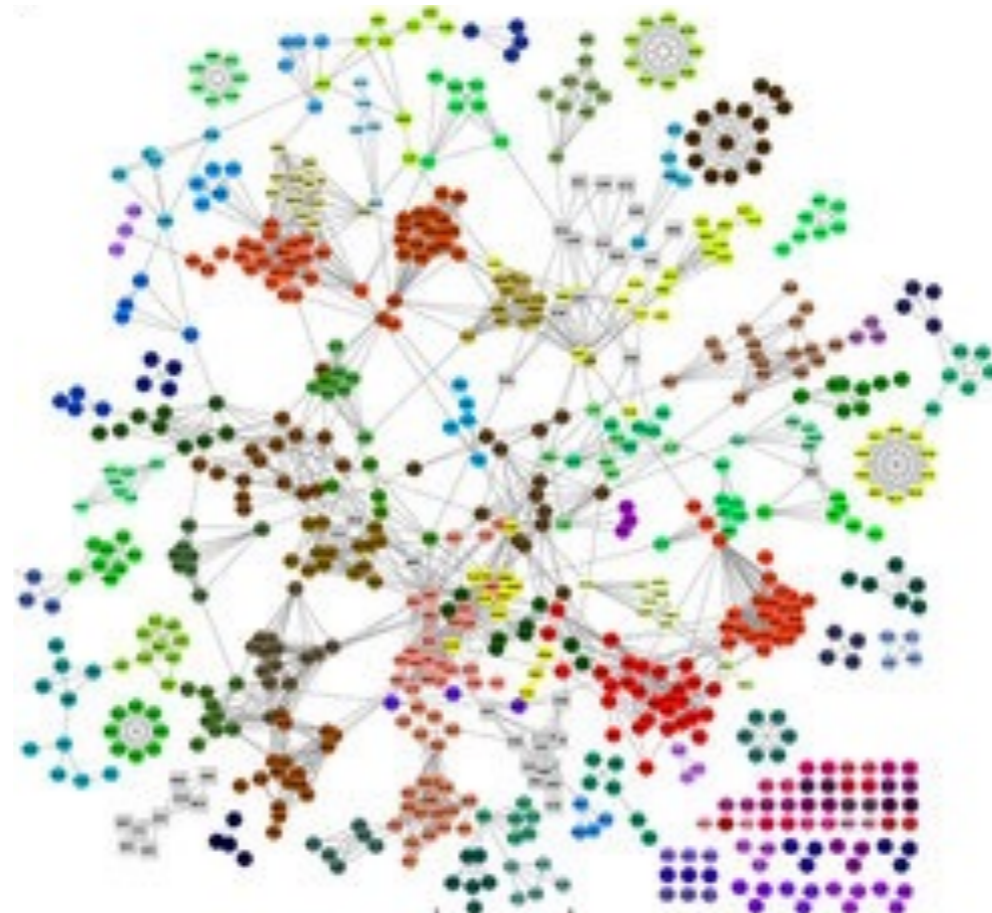
Random Graph Generation

- **Random Graph:** Given a graph with N nodes and all the possible $N \times (N-1)/2$ undirected edges select each of them with an independent probability P
- **Small-World Graph:** Start with a N nodes, generate a regular graph connect each of them with $k/2$ neighbors on each side. Rewire each of them with a probability P .
- **Scale-Free Graph:** Network of N nodes and kN edges. Start with $k+1$ nodes. Add one node at the time connecting k edges to the previous nodes with probability proportional to the number of edges of each of them (see example with $k=2$).



Community or Cluster

One of the main feature of the biological network is the **presence of subset of nodes densely interconnected (communities or clusters)**



Gaiter, Scientific Reports 2015

Clusters are important to detect similarity between nodes (genes, diseases, etc) in the same cluster.

Network Robustness

Robustness, the ability to withstand failures and perturbations. It is a critical attribute of many complex systems including biological networks.

Robustness is tested **removing nodes** and checking if **connections** between the remaining nodes **are conserved**. This is possible because may exist alternative paths between two distinct nodes.

Biological networks persists despite the environmental noise, mutations etc.

Telecommunication networks resist to the attack of hackers and hardware failure

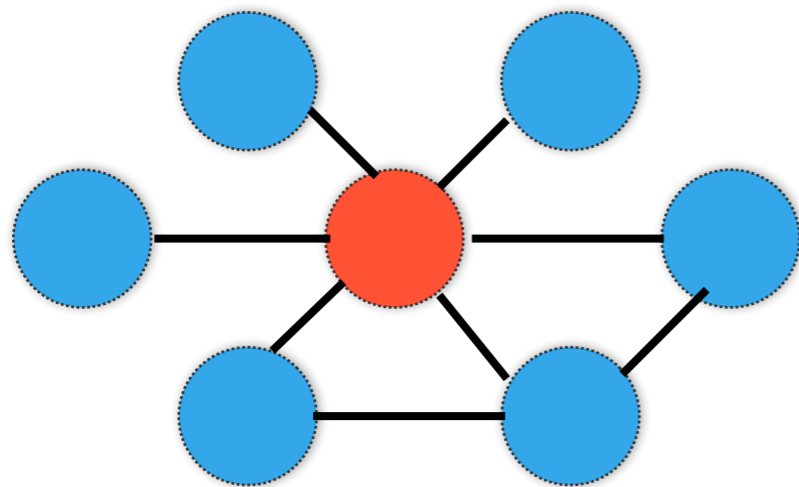
Network Perturbation

For **random networks the effect** of removing a single node is **on average the same**.

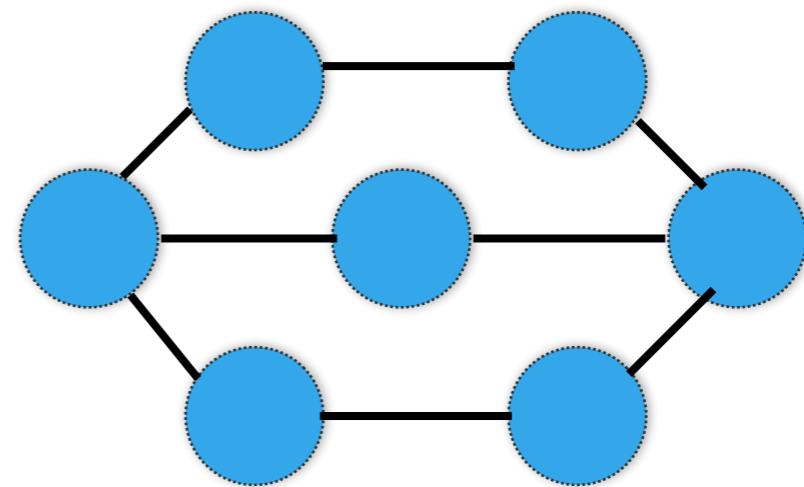
Biological networks are characterized by a small fraction of nodes with high degree (**hubs**)

An **attack that aims to a hub has strong effect** on the connectivity of the network.

In normal situation we assume that **attacks are random**. Thus, on average, an attack should have smaller effect on Biological Network.



Biological network

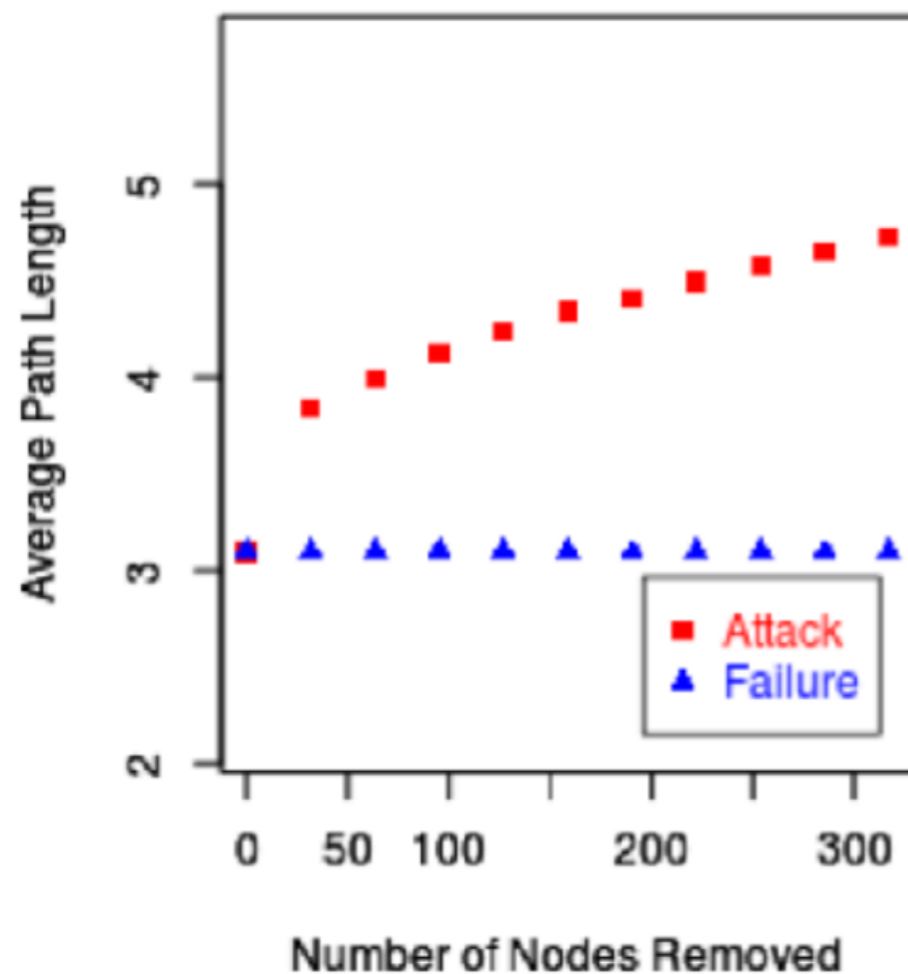


Random network

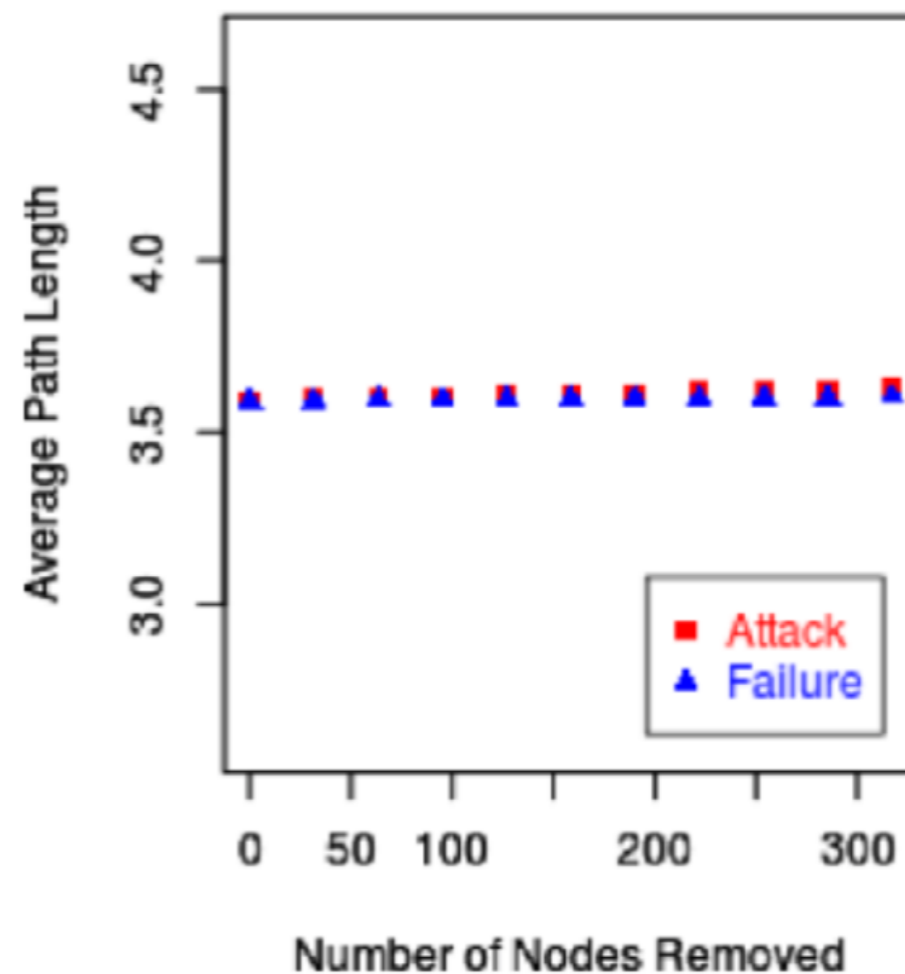
Failures and Attacks

The **resilience of scale free networks to failures** comes at the price of high **vulnerability to targeted attacks**

Homo Sapiens



Random Network



Python NetworkX

NetworkX is a Python package for the **creation, manipulation, and study** of the structure, dynamics, and functions of **complex networks**.

```
>>> import networkx as nx
```

```
>>> G = nx.Graph()
```

```
>>> G.add_node(1)
```

```
>>> G.add_nodes_from([2, 3]) # add list of nodes
```

```
>>> G.add_edge(1, 2)
```

```
>>> G.add_edges_from([(1, 2), (1, 3)]) # add list of edges
```

```
>>> G.number_of_nodes()
```

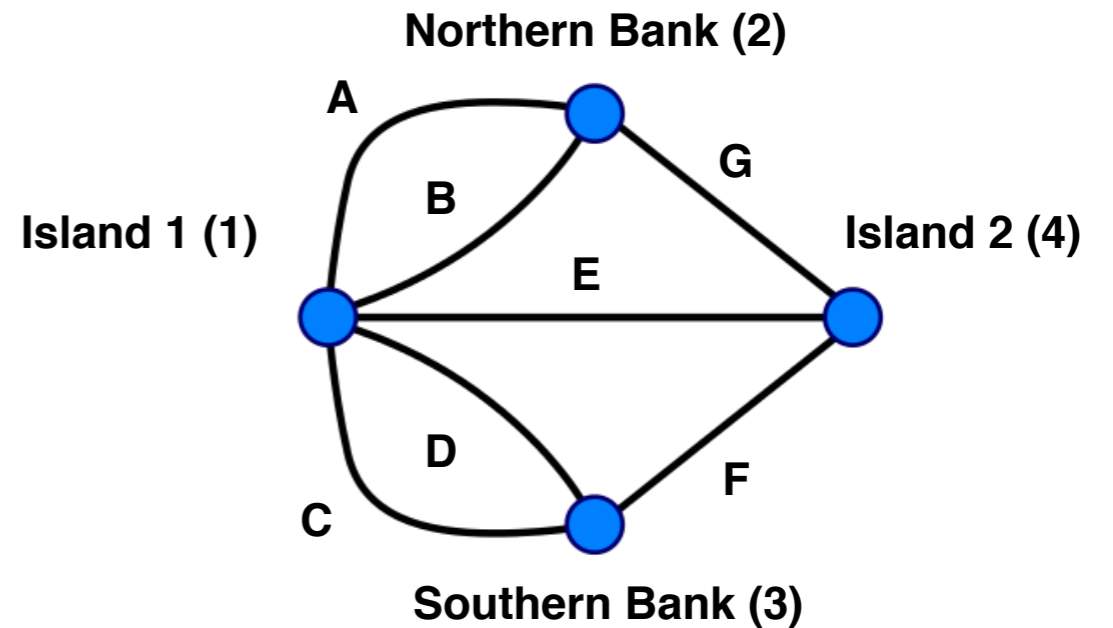
```
3
```

```
>>> G.number_of_edges()
```

```
2
```

Königsberg Graph

NetworkX allows to create graphs with **multiple edges connecting the same nodes** using the MultiGraph object.



```
>>> import networkx as nx
>>> M = nx.MultiGraph()

>>> M.add_edges_from([(1, 2, {"name": "A"}),
...                  (1, 2, {"name": "B"}), (1, 3, {"name": "C"}),
...                  (1, 3, {"name": "D"}), (1, 4, {"name": "E"}),
...                  (3, 4, {"name": "F"}), (2, 4, {"name": "G"})])

>>> M = M.degree(1)
```

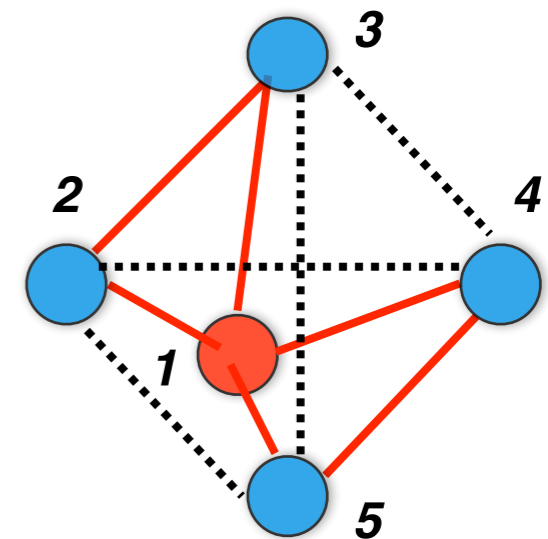
Exercise

- Calculate the **clustering coefficient of node 1**

$c(1)$ = the number of triangles with fix vertex in node 1 divided by the all possible triangles.

triangles = [(1,2,3),(1,4,5)]

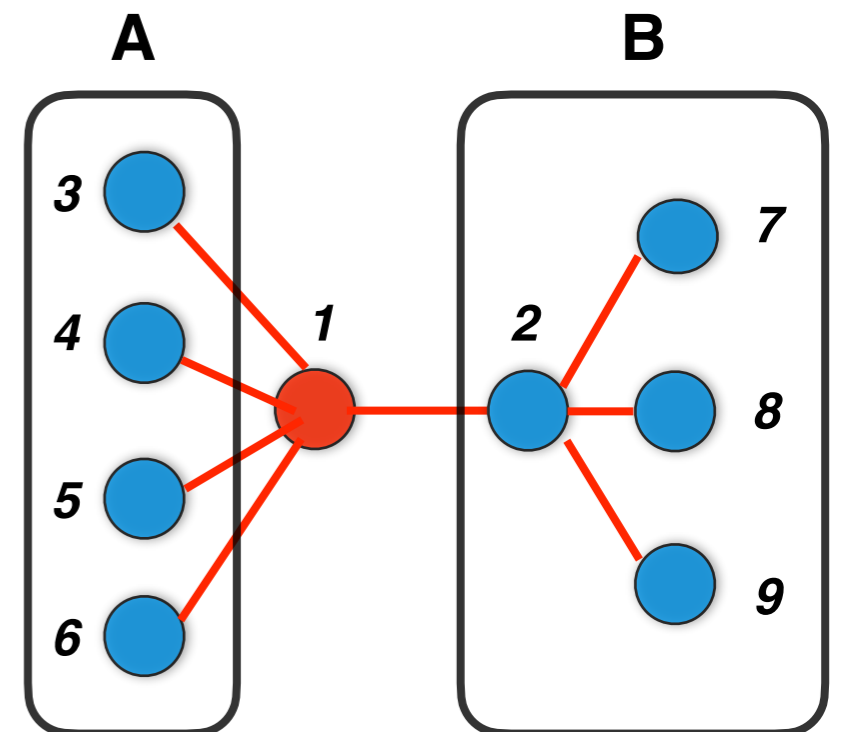
other triangles = [(1,3,4), (1,2,5), (1,2,4), (1,3,5)]



- Calculate the **number of shortest paths passing through node 1**

$\sigma(1)$ = all the paths from block A to block B
+ all the possible paths between nodes within block A

3 → 2, 7, 8, 9	3 → 4, 5, 6
4 → 2, 7, 8, 9	4 → 5, 6
5 → 2, 7, 8, 9	5 → 6
6 → 2, 7, 8, 9	



Network generators

Networkx has function that generate standard network types

```
>>> import networkx as nx
>>> import matplotlib.pyplot as plt

>>> er = nx.erdos_renyi_graph(100, 0.15)
>>> ws = nx.watts_strogatz_graph(30, 3, 0.1)
>>> ba = nx.barabasi_albert_graph(100, 5)

>>> nx.draw(ba)
>>> plt.show()
```


BioGRID

The **Biological General Repository for Interaction Datasets** (BioGRID) is a curated biological database of protein-protein interactions, genetic interactions, chemical interactions, and post-translational modifications

The screenshot shows the BioGRID website homepage. At the top left is the BioGRID 4.2 logo. A navigation bar contains links for home, help, wiki, projects, tools, contribute, stats, downloads, partners, and about us, along with a Twitter icon. The main content area features a welcome message: "Welcome to our Database of Protein, Genetic and Chemical Interactions". Below this is a paragraph describing the database's scope: "BioGRID is a biomedical interaction repository with data compiled through comprehensive curation efforts. Our current index is version 4.2.191 and searches 75,760 publications for 1,992,321 protein and genetic interactions, 29,093 chemical interactions and 959,750 post translational modifications from major model organism species. All data are freely provided via our search index and available for download in many standardized formats." Below the text are two buttons: "BioGRID Statistics" with a bar chart icon and "Latest Downloads" with a download icon. To the right is a search box titled "Search BioGRID:" with a dropdown menu set to "By Protein/Gene". The search input field contains the placeholder text "Search by Protein/Gene Identifiers ...". Below the search field is a dropdown menu for "All Organisms" and a "Submit Identifier Search Q" button. At the bottom of the search box are three icons: a magnifying glass for "Advanced Search", a flag for "Helpful Search Tips", and a star for "Featured Datasets". Below the search box is a yellow banner for the "BioGRID COVID-19 Coronavirus Curation Project" with links to "Search BioGRID for SARS-CoV-2 Protein Interactions" and "Download SARS-CoV-2 and Coronavirus-Related Interactions". The "Related Resources" section includes "BioGRID Themed Curation Projects" with a description and a "Learn more" link, and a "Partners" section with logos for NIH, ORIP (Office of Research Infrastructure Programs), CIHR IRSC, Mount Sinai Hospital, Princeton University, and Université de Montréal.

Exercise

Generate the three types of network (random, "small world" and "scale free") and calculate the distribution of the degree, betweenness and clustering.

From BioGRID download the Yeast interactome and analyze it with networkx importing only a list of unique interactions from the BioGRID repository:

- How many components are present?
- What are its/their number of nodes and edges?
- What is the the average values of degrees, clustering and betweenness?
- What is the gene with highest degree?
- Draw the distribution of the degree and fit to a power law distribution.