

# Chapter 1

## Methodology

### 1.1 Complex networks

The complex system has several properties. It consists of many components such as units or individuals who interact with each other. The properties of the complex system can not be predicted from the behaviour of one individual. In such systems, without any central force, collective behaviour can emerge. In societies, people's interactions lead to civilisation, economy, formation of social groups or even traffic on the roads. In the animal, populations are present at different levels of the organisation, as in ants and bees colonies or the school of the fishies showing flocking patterns. [?]

The research in complex systems focuses on the structure of the interactions between units. Knowing how branches of the system are connected, we can determine the emergence of the collective behaviour of the system. We can construct networks with neurons and synapses, representing their connections. The structure of the brain network and its properties are fundamental for brain functioning, and neurons in the same brain area are closely connected. Similarly, we can represent the communication between people. The structure of these interactions gives us insights, for example, how information propagates through the system. The presence of people with many connections can lead to faster information flow.

Despite the differences between complex systems, they can be studied using complex networks; with sets of nodes and edges. Elements in the system are nodes, while interactions between them are given as edges. This approximation allows us to treat equally social (graph of actors), biological (network of proteins) or even technological systems (internet, traffic). In recent years, complex network theory has application in different fields, and the availability of big data incurs its development.

The complex network theory originates from the graph theory in mathematics. These days, the graph and network are used as equivalent terms. The first mathematical problem solved using graph theory was *Konigsberg* problem of seven bridges. The city *Konigsberg* had seven bridges connecting the city's parts across the river and the island in the middle. The question was, is it possible to find a walk that crosses all seven bridges only once. Representing the problem as a graph, Euler managed to simplify the problem; the parts of the land are represented as nodes while bridges between them are links. Crossing each bridge only once is possible if each part of the land has an even number of connections. Thus, it was not possible in this case, as each piece of land had an odd number of bridge connections, see Fig. 1.1.

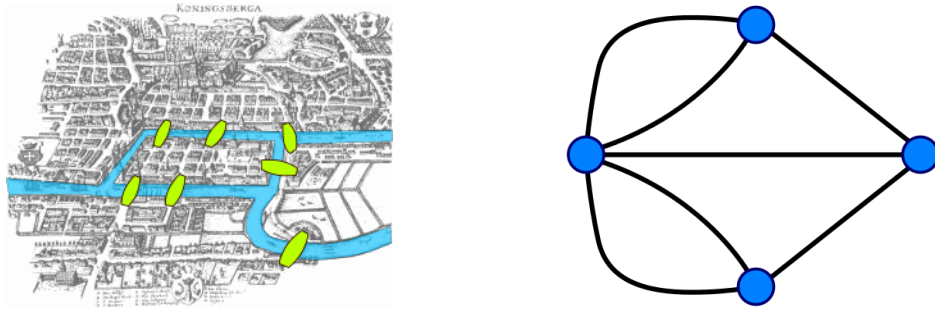


Figure 1.1: The Kronigsber problem of seven bridges.

## 1.2 Types of networks

The graph or network  $G$  is defined as  $G = (V, E)$ , where  $V$  is a set of nodes (vertices), and  $E$  is a set of edges. The edge is pair of nodes  $e_{ij} = (i, j)$ , and  $i, j \in V$ . The graph structure has undirected edges meaning that edges are symmetric:  $(i, j)$  implies  $(j, i)$ . Edges are also unweighted, meaning that all edges are equally important. The specific properties of the nodes are also neglected in this representation. The **adjacency matrix**  $A = N \times N$  has value 1 if there is connection between two nodes, otherwise it is 0 [?]

For example, if we consider unweighted, undirected network, with only 3 nodes and two connections, adjacency matrix is:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \quad (1.1)$$

The self loops usually are not considered, meaning that  $A_{ii} = 0$ .

Other equivalent way of representing graph is as edge list. Instead of having adjacency matrix, graph is described with the list of links that are in graph. The graph is then pair of  $N, g$ , where  $N$  is set of nodes, and  $g$  is collection of links, listed as subset of  $N$  of size 2. So the network is written as  $g = \{\{1, 2\}, \{2, 3\}\}$ .

Sometimes it is essential to include specific properties of the system in the network representation, which can help create more realistic models. The additional properties can be added on the edge, node or network level.

In a **directed network**, edges have broken symmetry. The interaction from node  $i$  to node  $j$  does not need to have the same property as the interaction from node  $j$  to node  $i$ . A typical example is WWW, where webpages are nodes, and hyperlinks are directed edges. In biological networks, gene regulation and neural activation can be given as directed networks.

The frequency of interaction between nodes is emphasised if edges are associated with different scalar values; networks are **weighted**. The edges may be signed, representing activation in the biological system or trust and distrust in the social system. In general, edges can be associated with any categorical variable. If attribute describes the time when an interaction between nodes happened; network is called **temporal**. Finally, **multigraph** allows the presence of multiply edges between two nodes. For the network between cities, edges may be different driving paths between them. In neuron cells, multiply synapses are represented as distinct edges.

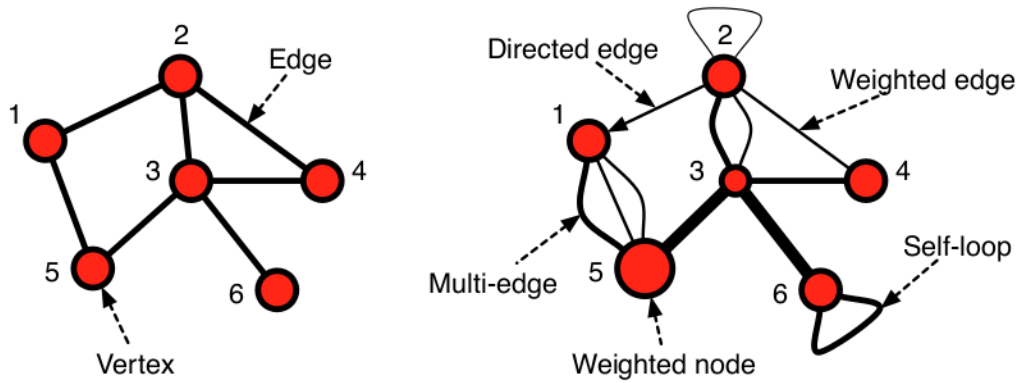


Figure 1.2: Graph types

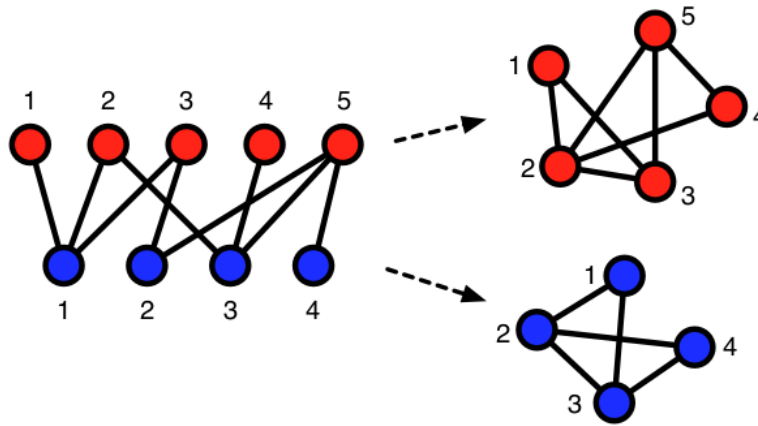


Figure 1.3: Graph types

A **bipartite network** has two partitions,  $U$  and  $V$ . The nodes in the same partition are not connected while links exist only between nodes of a different kind. In general, we can define  $k$ -bipartite graph. The set of nodes  $V$  has  $k$  distinct classes of nodes. When  $k = 2$  the network is bipartite. Bipartite networks represent the membership of people or items in groups. For example, we can define the network of actors as a bipartite graph. In one partition are actors and in other movies. There are no edges between actors or movies, but the actor is connected to the film if it plays in that movie. Another example is a recommender network, such as a network of people and items they like.

The equivalent representation of bipartite network is incidence matrix  $B$ . If  $n$  is number of people and  $g$  number of groups, this matrix is  $g \times n$ , having elements  $B_{ij}$  1 if person  $i$  belongs to group  $j$ .

Even bipartite networks give realistic representation of the system, there is often need to analyze the single type of nodes. From a bipartite network, we can generate two projections. The first one connects nodes partition  $V$  if they point to node  $u$ . Similarly, we can project the network on  $U$  partition, connecting  $u$  nodes. The one mode projection between actors and movies onto actors is undirected network of actor collaborations. Actors are connected if they appear in the same movie. We can also create one-mode projection onto movies, where two movies are connected if they share the same actor.

The projections are useful in some manner, but they also lose some important information, for example how many groups nodes share in common. This information can be propagated adding the weight to the edges, equal to the number of common groups.

The product  $B_{ki}$  and  $B_{kj}$  is 1 if  $i$  and  $j$  belong to the same group  $k$ . Thus the total number of groups to which nodes  $i$  and  $j$  belong is  $P_{ij} = \sum_{k=1}^g B_{ki}B_{kj} = \sum_{k=1}^g B_{ik}^T B_{kj}$ . The matrix  $P$  is matrix of one-mode projection. The diagonal elements are non-zero, and represent the number of groups node  $i$  belongs to. To derive the weighted adjacency matrix, the diagonal elements are set to 0. The adjacency matrix of unweighted projection, each non-zero element needs to be replaced with 1.

The important consequence of the one-mode projection is the construction of the cliques; subgraph in which every pair of nodes is connected. Every node that is being projected is represented as a clique of size, because all pairs of its neighbours are exactly distance two away from each other. All actors in the movie will be joined in a clique in the one-mode actor projection.

Another consequence of the one-mode projection is that one mode projection may originate from different bipartite networks. Meaning that projection is not one-to-one; no bijection. It is surjective operation, meaning hat any projected network  $P$  has at least one bipartite network such that the projection of  $B$  results in  $P$ .

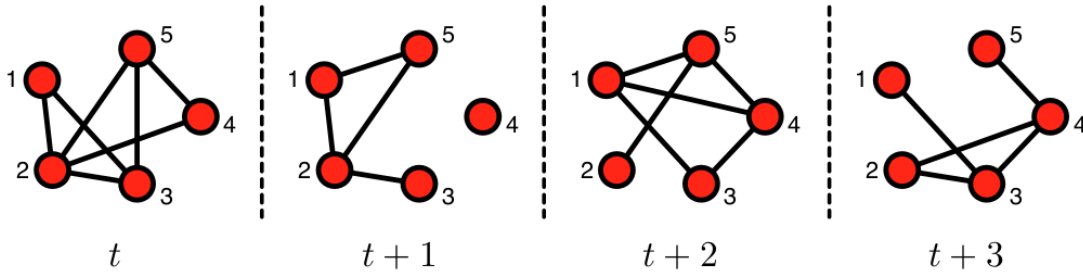


Figure 1.4: Graph types

In **temporal networks** nodes and edges evolve. Many real networks are not static, networks grow over time, and edges and nodes may emerge or disappear. Also, some edges may be active at regular intervals, reflecting the circadian rhythms of the nodes. In citation networks, new nodes (paper) join the network, creating links with cited papers.

Let consider the temporal network with  $N$  nodes, over time interval  $t_{max}$ . The event representation consists in viewing the temporal networks as a collection of time-stamped edges. In this representation, each edge  $(i, j)$  is defined as  $(i, j, t, \Delta t)$ , where  $t$  is the time of the event and  $\Delta t$  its duration. In a temporal network, the same link can appear multiple times, and the duration of the event may vary. An example of an event temporal network may be phone-calls networks, where a call between two persons  $i$  and  $j$  started at time  $t$  and ended at  $t + \Delta t$ .

A temporal network can be represented as sequence of networks  $G = G(1), \dots, G(t_{max})$ , where  $t_{max}$  is number of networks. While in the event-based representation, time can be both discrete and continuous, in the snapshot representation, time is only discrete. The temporal network is seen as a structure that evolves in time, and at each timestamp, we can analyse the system's macroscopic properties. We need to specify time windows for coarse-grain event-based representation of the temporal information in the data. If we use uniformly

time window of length  $w$  then the events occurring in  $0 \leq t < w$  enter snapshot  $G(1)$ , those occurring in  $w \leq t < 2w$  enter  $G(2)$  and so forth. We can not recover the original data points from the snapshot representation, and more information loss is present with a larger time window. If the time window is  $w = t_{max}$ , there is only one snapshot, temporal data are no more available, and the network is static.

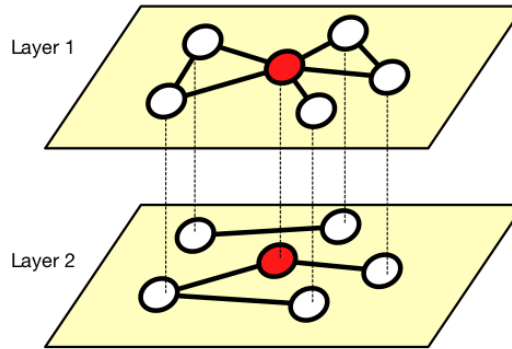


Figure 1.5: Graph types

A network in which edges are marked by which "layer" they exist in is called a multiplex or multilayer network. These networks are used to represent a system in which there are multiple types of interactions, and we store the connectivity of each kind in a different "layer" of the multiplex network. A temporal network is a special kind of multiplex network, where these layers form a temporal (ordered) sequence. Crucially, there can dynamics on each vertex that govern which layer some kind of interaction occurs on, so multiplex networks are not merely a special kind of graph in which different colors or layer numbers annotate edges.

Spatial networks are a special kind of node-annotated network, in which the annotations represent the node's location in some  $d$ -dimensional space. This graph property is most common in transportation networks, e.g., as road and city networks, airport transportation networks, oil and gas distribution networks, shipping networks, etc., but can also appear in social networks. Planar graphs are a special case of spatial networks in which the nodes are embedded on a 2-dimensional surface and edges do not cross.

Hypergraphs are another type of network, in which edges denote the interaction of more than two vertices, e.g,  $E$  in  $V \times V \times V$ . Scientific collaboration graphs can be represented as a hypergraph, in which each "edge" is the set of coauthors on a scientific article. However, collaboration networks are more commonly represented as bipartite graphs, in which scientists and papers form two sets of vertices, and scientist-nodes are connected to all the paper-nodes on which they are authors.

## 1.3 The structure of complex networks

### Paths and cycles

As network is connected structure, the nodes may be influenced also by distant nodes, and the information might spread through the links of a network. Analyzing the paths of the network is important task.

A path of the network between two nodes  $i$  and  $j$  is a sequence of links,  $i_1 i_2, \dots, i_k$  such  $i_k i_{k+1} \in g$  that  $i_1 = i$  and  $i_k = j$ , each node in the sequence is distinct. The path is walk, if

nodes in the sequence are not distinct, so walk can visit one node more than once. The cycle is walk that starts and ends at the same node, while other nodes are distinct.

Note that, if we use adjacency matrix representation where  $A_{ii} = 0$ , then  $A^2$  tells us how many walks of length 2 exist between any two nodes.

Network is connected if every two nodes in the network are connected by some path. So the network is connected if for every node  $i \in N$  and node  $j \in N$  exists a path between them.

A component of the network are distinct maximal connected subgraphs of a network. In the example there are 4 components. Note that in this definition of the component isolated node is also component. In directed network, set of nodes that are reachable from each other is a strongly connected component, while a set of nodes where either  $i$  is reachable from  $j$  or  $j$  is reachable from  $i$  is weakly connected component.

There are a few particular network structures that are commonly referred to. A tree is a connected network that has no cycles. A forest is a network such that each component is a tree. Thus any network that has no cycles is a forest, as in the example pictured in Figure 2.1.6. A particularly prominent forest network is a star. A star is a network such that there exists some node  $i$  such that every link in the network involves node  $i$ . In this case  $i$  is referred to as the center of the star. There are a few facts about trees that are easy to derive (see Exercise 2.2) and worth mentioning. A connected network is a tree if and only if it has  $n - 1$  links. A tree has at least two leaves, where leaves are nodes that have exactly one link. In a tree, there is a unique path between any two nodes. The complete network is one where all possible links are present, so one where  $g_{ij} = 1$  for all  $i \neq j$ .

A circle (also known as a cycle-graph) is a network that has a single cycle and such that each node in the network has exactly two neighbors. In the case of directed networks, there can be many different stars involving the same set of nodes and having the same center, depending on which directed links are present between any two linked nodes. On occasion, it will be useful to distinguish between these.

### Eulerian tours and Hamiltonian cycles

A walk is said to be closed if it starts and ends at the same node. It is clear that in order to have a closed walk that involves every link of a network exactly once it must be that each node in the network has an even degree. This follows since each time a node is “entered” by one link on the walk it must be “exited” by a different link, and each time the node is visited, it must be by a link that has not appeared previously on the walk. Euler’s [?] simple but remarkable theorem is that this condition is necessary and sufficient for there to exist such a closed walk.

A connected network  $g$  has a closed walk that involves each link exactly once if and only if the degree of each node is even.

One can ask a related question for nodes rather than links: when is it possible to find a closed walk that involves each node in the network exactly once? Such a closed walk must be a cycle, and is referred to as a Hamilton Cycle or a Hamiltonian. A related question is whether there exists a “Hamilton path” that hits each node exactly once. Clearly a network that has a Hamilton cycle has a Hamilton path, while the converse is not true (consider a line). Discovering whether or not a network has a Hamilton cycle is a much more challenging question than whether it has a Euler tour; and this has been an active area of research in graph theory for some time. It has direct applications to the “traveling salesman problem,” where a salesman must visit each city on a trip exactly once, cities are nodes on a network, and the path must follow the links. The seminal theorem on Hamilton cycles is due to Dirac [?]. Stronger theorems have since been developed, as we shall shortly see, but it is worth stating on its own, as it has an intuitive proof that helps one see the paths to proving some of the later.

## Community structure

## Core-periphery structure

# 1.4 The network analysis

## Degree distribution

The simplest network measure is **node degree**,  $k$ . The degree of node  $i$  gives the number of nodes attached to node  $i$ ,  $k_i = \sum_j A_{ij}$ .

The density of the network is average degree divided by  $N - 1$ , where  $N$  is number of nodes. It is relative fraction of nodes in the network.

In the case of regular networks, such as grids, each node has an equal degree, meaning that nodes in the network have similar roles. In the general case, the networks have more complicated structure. If degree sequence is skewed, we are able to identify nodes with high degree, hubs. Removing hubs may partition a connected network into several components. Finally, if we are able to test isomorphism between two graphs, the starting point would be to compare their degree sequences are the same. If they are not, then they can not be isomorphic.

To calculate the degree distribution we can consider the fraction of  $k$  degree nodes  $N_k$ ,  $p(k) = N_k / N$ . It is the probability,  $P(k)$ , that randomly chosen node has degree  $k$ . Similarly, we can order nodes according to their degree and plot the node degree.

If the nodes of the graph are statistically independent, the degree distribution completely determines the properties of a network. If degree distribution is present  $P(k)$ , and the total number of nodes is fixed to  $N$ . We can construct a graph with random connections. Label  $N$  vertices. To node  $j$  of the graph ascribe degrees  $k_j$ , taken from the distribution  $P(k)$ . Connect at random ends of pairs of distinct quills belonging to distinct vertices.

- The Poisson distribution. The degree distribution in random network, where all nodes have the same connecting probability, follows Poisson distribution  $P(k) = \frac{(Np)^k e^{-Np}}{k!}$ , where  $k$  is the mean degree distribution.
- Exponential distribution.  $P(k) = e^{-k/k}$ . This is degree distribution of the growing random graph. Even for infinite networks all moments of distributions are finite, and have natural scale of the order of average degree.
- In real networks degree distribution follows a power law.  $P(k) = k^{-\gamma}$ , where  $\gamma$  is exponent of the distribution. In this distribution there is no natural scale, so they are called scale-free networks. In infinite networks all higher moments diverge. If the average degree of scale-free networks is finite, than  $\gamma$  exponent should be greater than 2. Therefore, real networks have a scale-free structure with the emergence of the hubs [?]. In finite size networks, fat-tailed degree distributions have natural cut-offs.

When plotting the degree distribution, it is common to use scaling of the axis. As explained, we rank the nodes according their degree and plot the node degree of each  $k$ th node. In the first example, we used linear scale for both axes. As many nodes have low degree, it is more useful to use logarithmic scale. On the logarithmic scale the distance between two points is proportional to the logarithm of distance between two points, meaning that distance on logarithm scale is same between points 10 and 100, and 100 and 1000. Now it is more easily notices that data-points follow straight line, meaning that degree distribution is some kind of exponential function.

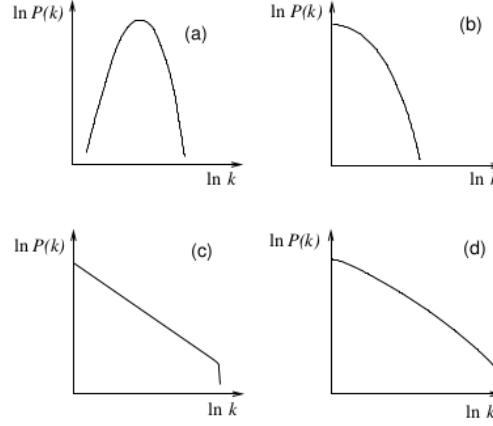


Figure 1.6: Distributions, ovde hocemo 3 distribucije+ linearna vs. logarithm scala, takodje pokazati kako izgleda frequency order

## Degree correlations

Correlation is defined through a correlation coefficient  $r$ . If  $x$  and  $y$  are two stochastic variables, for which we have a series of observation pairs  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . The correlation coefficient  $r(x, y)$  between  $x$  and  $y$  is defined as:

$$r(x, y) = \frac{\frac{1}{n} \sum_{i=1}^n ((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (1.2)$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ , is the average over variable  $x$ .

Taking the definition of correlation coefficient we can define it for vertex degrees. For simple graph  $G$  with vertex set  $V(G) = \{v_1, \dots, v_n\}$ ,  $A[i, j] = 1$  if there is a link between nodes  $v_i$  and  $v_j$ . If  $G$  is a simple graph with adjacency matrix  $A$  and degree sequence  $\mathbf{d} = [d_1, \dots, d_n]$

$$r_{deg}(G) = \frac{\sum_{i=1}^n \sum_{j=1}^n ((d_i - \bar{d})(d_j - \bar{d})A[i, j])}{\sum_{i=1}^n (d_i - \bar{d})^2} \quad (1.3)$$

Using adjacency matrix, allow us to calculate the correlations between neighboring nodes. If two nodes are not connected  $A[i, j] = 0$ , the degree correlation between them do not have contribution to the  $r$ .

The **degree-degree correlations** in the network are measured by **assortativity**. If correlations are positive, networks are assortative; there is a tendency that connections exist between similar degree nodes. The negative correlations indicate that large degree nodes have preference to connect nodes with small degree; disassortative networks. The average first neighbor degree  $k_{nn}$  can be calculated as  $k_{nn} = \sum_{k'} k' P(k'|k)$ . The  $P$  is conditional probability that an edge of degree  $k$  points to node with degree  $k'$ . The norm is  $\sum_{k'} P(k'|k) = 1$ , and detailed balance conditions  $[?], kP(k'|k)P(k) = k'P(k|k')P(k')$   $[?]$ . If the node degrees are uncorrelated,  $k_{nn}$  does not depend on the degree, otherwise increasing/decreasing function indicates on positive/negative correlations in the network.

The Newman defined the assortativity index  $r$  in slightly different way:



$$r = \sum_{kl} kl(e_{kl} - q_l q_k) / \sigma_q^2 \quad (1.4)$$

where  $e_{kl}$  is the probability that randomly selected link connect nodes with degrees  $k$  and  $l$ ,  $q_k$  is probability that randomly chosen node is connected to node  $k$  and equals  $q_k = kp_k / \langle k \rangle$ , while  $\sigma_q$  is variance of the distribution  $q_k$ .

## Distances

Between two nodes in the network, we can define different paths, but the most important one are the shortest paths. The distance between two nodes  $d(i, j)$  is defined as the length of shortest path between two nodes. In the case of weighted networks, is defined such that the path has minimal weight, and the length of such path does not have to be minimal. Distances on the network, again can give us insight how to networks are similar, and to give indication of relative importance of the node in the network.

The eccentricity of the vertex, gives us how far the farthest vertex is positioned in the network. The radius is minimum over all eccentricity values, while the diameter defines the largest distance between nodes in the network. These definitions apply to directed and undirected graphs.

The diameter gives us useful information, it may not be powerful enough to discriminate among graphs. An important metric is to consider the distribution of path lengths. The average distance between nodes is useful for network description.

If  $G$  is connected graph with vertex set  $V$  and  $\bar{d}(u)$  is average length of the shortest paths from node  $u$ , to any other node  $v$  in network  $G$ .

$$\bar{d}(u) = \frac{1}{|V| - 1} \sum_{v \in V, v \neq u} d(u, v) \quad (1.5)$$

From there we can define the average path length as mean value over  $\bar{d}(u)$ .

$$\bar{d}(G) = \frac{1}{|V|} \sum_{u \in V} \bar{d}(u) \quad (1.6)$$

while the characteristic path length of  $G$  is median over all  $\bar{d}(u)$ .

## D-measure

For each node  $i$  we can define the distribution of the shortest paths between node  $i$  and all others nodes in the network,  $P_i = \{p_i(j)\}$ , where  $p_i(j)$  is percent of nodes at distance  $j$  from node  $i$ . The connectivity patterns can efficiently describe difference between two networks. To specify how much  $G$  and  $G'$  are similar we use D-measure [?]

$$D(G, G') = \omega \left| \sqrt{\frac{J(P_1, \dots, P_N)}{\log(d)}} - \sqrt{\frac{J(P'_1, \dots, P'_N)}{\log(d')}} \right| + (1 - \omega) \sqrt{\frac{J(\mu_G, \mu_{G'})}{\log 2}} \quad (1.7)$$

D-measure calculates Jensen-Shannon divergence between  $N$  shortest path distributions,

$$J(P_1, \dots, P_N) = \sum_{i,j} p_i(j) \log\left(\frac{p_i(j)}{\mu_j}\right) \quad (1.8)$$

where  $\mu_j = (\sum_{i=1}^N p_i(j)) / N$  is mean shortest path distribution.

The first term in equation 1.7 compares local differences between two networks, and Jensen-Shannon divergence between  $N$  shortest path distributions  $J(P_1, \dots, P_N)$  is normed with network diameter  $d(G)$ . The second part determines global differences, computing  $J(\mu_G, \mu_{G'})$  between mean shortest path distributions. The D-measure ranges from 0 to 1. The lower D-measure is, networks are more similar and for D-measure  $D = 0$ , structures are isomorphic.

### Clustering coefficient

The **clustering coefficient** is a measure describing the neighbourhood's structure. In networks exist tendency to form triangles or clusters. This is common in friendship networks where two friends of one person have a high probability of being friends. The clustering can be measured by computing the number of links between neighbours of one node,

$$c_i = 2e_i / (k_i(k_i - 1)) \quad (1.9)$$

Averaging it over all network nodes, we can calculate the mean clustering coefficient. It ranges from  $\langle c \rangle = 0$  where connections between neighbouring nodes do not exist, network has the structure of three. On the other hand,  $\langle c \rangle = 1$  indicates a fully connected network.

Newman proposed the alternative definition for the clustering coefficient based on the number of triples and triangles in a graph. A triangle at node  $v$  is complete subgraph with 3 nodes, including  $v$ . A triple on the node  $v$  is a subgraph of exactly three nodes and two edges, where  $v$  is incident with two edges. The network transitivity is defined as the ratio of number of triangles in the network over the number of triples. The network transitivity is seen as global clustering, as it considers the whole network.

### Real-world networks

Real-world networks share similar properties. The mean distance between nodes is smaller than the number of nodes in the network  $l \ll N$ , called small-world phenomena. This cause the fast spread of information or even diseases in the complex systems. In small-world networks number of vertices grow exponentially with distance; thus  $l$  increase as  $\log(n)$  or slower. Logarithmic scaling can be proved from various network models; also, it is observed in real-world complex systems. The clustering coefficient in real-world networks is usually high. Real-world networks have one important feature; power-law degree distribution; such networks are called scale-free networks.

## 1.5 Network models

### Random network model

The random graph model was introduced by mathematicians Paul Erdős and Alfred Rényi in 1959. In this model, connections between nodes are chosen randomly, and every link has the same probability of existing. The graph is characterized only by a number of the nodes  $N$  and the linking probability  $p$ , so Erdős-Rényi graph is written as  $G(n, p)$ .

The creation of ER random network consists of the following steps:

- we start with  $N$  isolated nodes

- between each  $N(N - 1)/2$  pair of nodes we create link with probability  $p$ ; sampling random number  $r \in (0, 1)$ , we create link if  $r \leq p$

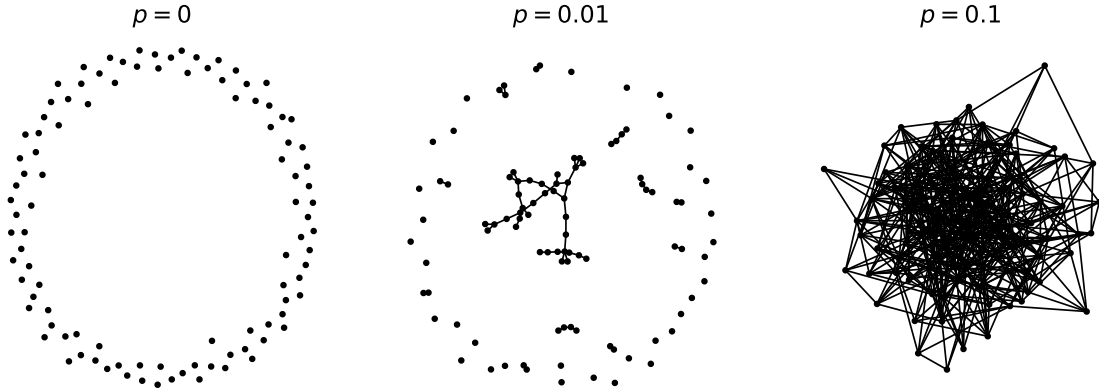


Figure 1.7: ER graph with  $N = 100$  nodes and different linking probabilities  $p$ .

We should note that this process is stochastic. The networks  $G(N, p)$  with the same parameters do not need to have the same structure; i.e. they differ in the number of links. Therefore, the single random graph is only one graph from all the possible realizations in the statistical ensemble.

Two simple quantities that could be estimated are the average number of links and the average degree. For complete graph with  $N$  nodes, number of edges is  $N(N - 1)/2$ . As the probability of drawing every edge is  $p$ , the **average number of links** is simply given as

$$\langle L \rangle = \frac{N(N - 1)}{2} p \quad (1.10)$$

From there, we conclude that the network's density is equal to probability  $p$ . The **average degree** is approximated as:  $\langle k \rangle = 2\langle L \rangle / N$ , leading to:

$$\langle k \rangle = (N - 1)p \quad (1.11)$$

The **degree distribution** of ER random graph follows the binomial distribution.

$$P(k) = \binom{N - 1}{k} p^k (1 - p)^{N - 1 - k} \quad (1.12)$$

The probability that the node has degree  $k$  is given with the second term  $p^k$ , while the probability that other  $N - 1 - k$  links are not created is given with the third part of the equation. Finally, there are  $\binom{N - 1}{k}$  combinations for one node, to have  $k$  links from  $N - 1$  possible links.

The binomial distribution describes very well small networks. For larger networks, we find that they are sparse and that the average degree is much smaller than a number of nodes  $\langle k \rangle \ll N$ . In this limit, binomial distribution becomes the Poisson, which now depends only on one parameter  $\langle k \rangle$

$$p(k) = \frac{1}{k!} e^{-\langle k \rangle} \langle k \rangle^k \quad (1.13)$$

The random graph has a very small **average path length**, it is given as  $\langle l \rangle = \frac{\ln N}{\ln(pN)}$  that is characteristic of many large networks. The clustering coefficient is proportional to linking probability,  $\langle C \rangle = p$ , so in large random networks, we find a small clustering coefficient,

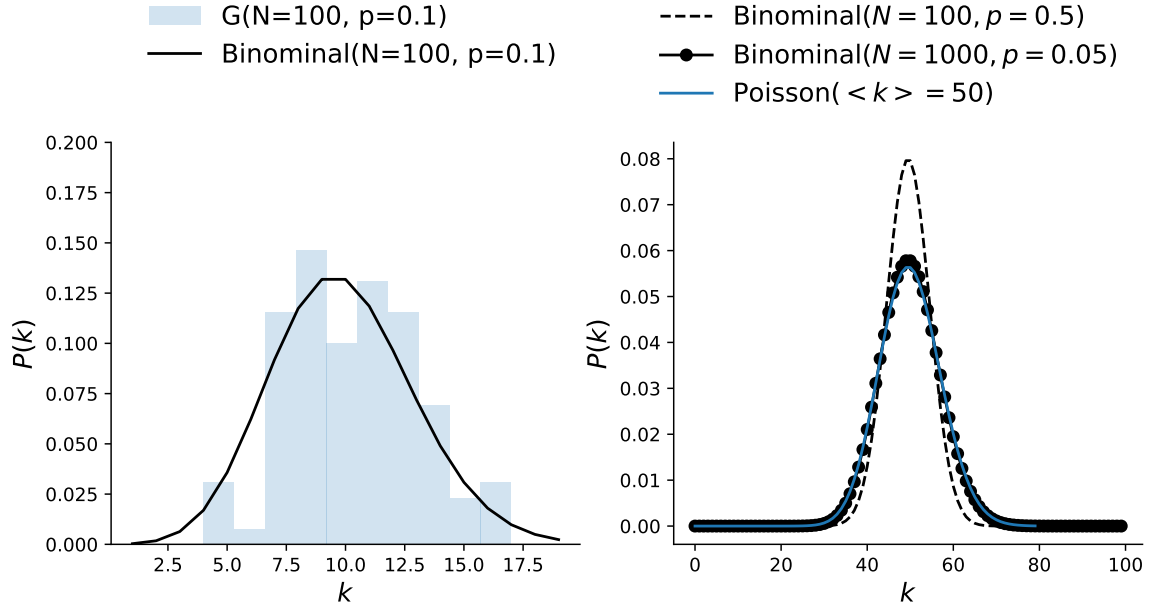


Figure 1.8: Degree distribution of ER graph. Degree distribution of small networks follow binominal. Larger networks are better approximated with Poisson distribution, and degree distribution for fixed average degree  $\langle k \rangle$  becomes independent of the network size.

contrary to real-world networks.

The figure 1.7 shows how the network becomes more connected by increasing the linking probability  $p$ . When  $p = 0$ , all nodes are disconnected. In the other limit,  $p = 1$ , the network is fully connected. Between those two probabilities exists critical probability, where the giant component appears. The giant component is a sub-graph, which size is proportional to the network size. In other words, the network does not have disconnected components. Such change in the network is a phase transition in network connectivity and is related to percolation theory.

The phase transition occurs when average degree is  $\langle k \rangle = 1$ , which gives us:  $p_c = \frac{1}{N-1}$ , meaning that all nodes have degree larger than 1. When the  $\langle k \rangle < 1$ , the network is in the sub-critical regime where all components are small. In the critical regime, the size of the giant component is proportional to the  $N^{2/3}$ . In the supercritical regime,  $\langle k \rangle > 1$ , the probability of a giant component appearing is 1.

## Small-world networks

Inspired by the idea that real-world networks are highly clustered and the average distance is small, Watts and Strogatz proposed the "small-world" model. The model starts from the regular lattice, and with rewiring links, the network starts to resemble small-world property. The procedure is the following:

- At the beginning, nodes are placed on the ring lattice, and each node is connected to  $k/2$  first neighbours on the left and the right side. Initially, the clustering coefficient is high,  $c = 3/4$ .

- For each link in the network, with probability  $p$ , we choose a random node to rewired the link. This makes long-distance nodes connect, decreasing the network's average path length.

The model interpolates between the regular graph when the probability is  $p = 0$  and the random graph with  $p = 1$  when all links are randomly rewired. Short distances and high clustering are present in the network for the critical probabilities.

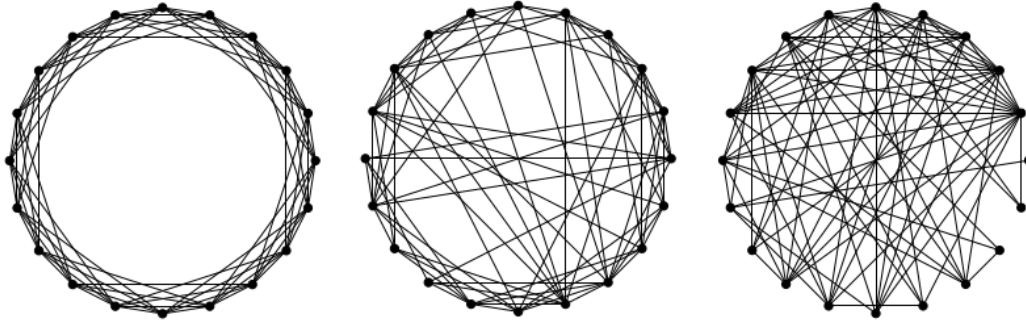


Figure 1.9: Watts and Strogatz graph model creation

Even though the small-world network model lacks the power-law degree distribution found in the real-world networks, it is an important model that motivated the research on random graphs.

### Barabási-Albert model

The ER random graph model and WS small-world model are static models, where the number of nodes is fixed. It is one of the reasons why they can not fully explain the properties of real systems. The size of real systems does not remain constant; real networks grow. For the network, the growth means that at each time step, new nodes are added to the network. The simplest model that produces the scale-free networks is Barabasi-Albert model.

- The model starts from the small number,  $n_0$  randomly connected nodes, with  $m_0$  links.
- At each time step, new node with  $m$  links joins to the network. New node creates links with the nodes already present in the network, following the linking rules; in this case rules of preferential attachment.

The preferential attachment is important ingredient for generating system with scale-free properties. In the real-system the linking between nodes is not random process, there exists the preference toward specific types of nodes. For example the popular web-pages can easily get more visits or it is common that already popular papers will get more citations. This effect is also called rich-get-richer or preferential attachment.

The simplest formulation of the preferential attachment model is that new nodes tend to connect with high degree nodes. The linking probability  $\Pi$  is then proportional to node degree  $k$ :

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \quad (1.14)$$

As at each time step one node arrive, we can estimate the number of nodes at the time step  $t$ ,  $N(t) = n_0 + t$ , with links  $L(t) = m_0 + mt$ .

First we can calculate the evolution of network degree in time.

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j} = m\frac{k_i}{m_0 + 2mt} \quad (1.15)$$

Note that new node, that arrived at time point  $t_i$  has degree  $m$ , as it links to  $m$  old nodes. Solving the equation we get that at  $t > t_i$ , has degree that grows as square root of time, also it shows that younger nodes easily acquire larger degree.

$$k_i(t) = m \left( \frac{t}{t_i} \right)^{\frac{1}{2}} \quad (1.16)$$

Degree distribution follows power-law, and for large  $k$  is aproximated with  $P(k) = k^{-\gamma}$ , such that  $\gamma = 3$ . More precisely, the degree distribution has form:

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \quad (1.17)$$

For large  $k$  it is exactly power-law. It is also independent of the time and size of the system, meaning the emergence of stationary scale-free state. Distributions do not depend on the  $N$ . If we vary  $m$  the slope of distributions is the same, but they are parallel. After re-scaling  $p(k)/m^2$ , they fall on the same line.

As network grows nodes with larger degree becomes bigger, so we end up with few nodes with many links, called hubs. The **network diameter**, represents the maximum distance in network,  $d \sim \frac{\ln N}{\ln \ln N}$ . The diameter grows slower than  $\ln N$ , making the distances in BA model smaller than in random graph. The difference is found for large  $N$ . Knowing that BA network has hubs, that shorten the path between less connected nodes. Also, if hubs are removed from the network, network easily partition in several components, loosing its properties.

The **clustering coefficient** of the BA model follows  $C \sim \frac{\ln N^2}{N}$ . It is different from clustering found in random networks, and BA networks are in general more clustered.

The combination of the growth and preferential attachment linking is crucial for getting scale free networks. For example, eliminating the preferential attachment; in growing network with random linking, degree distribution is stationary, but it follows exponential. In contrast, the absence of growth leads to the non-stationary degree distribution. When number of nodes is fixed, while the network grows only in number of links, such that randomly chosen node  $i$  connects to node  $j$  according to probability  $\Pi$ . At the beginning, the degree distribution follows the power-law, same as in BA model. As more links are added to the network, the distribution changes it's shape, first the peak appears, while at the end network becomes complete graph, where all nodes have the same degree.

### Nonlinear preferential attachment model

In the nonlinear preferential attachment model linking probability also depends on the node degree. The dependence is not linear and has the following a form:

$$\Pi(k_i) = k_i^\beta \quad (1.18)$$

The probability that newly added node attaches to node  $i$  depends on the existing  $i$ -th node degree  $k_i$ , and the parameter  $\beta$ . When  $\beta = 1$ , the model is BA model, where degree distribution follows the power-law. When  $\beta = 0$ , linking probability becomes uniform; i.e. it corresponds to random network model, and degree distribution is Poisson; there is exponential decay.

For  $\beta > 1$ , the effects of preferential attachment are increased, leading to emergence of super-hubs. The hub-and-spoke network appear in this regime, where almost all nodes are connected to few high-degree nodes.

On the other hand, if  $\beta < 1$ , the model is in so called sub-linear preferential attachment regime. The linking probability is not random so degree distribution does not follow Poisson; but also the preference toward high degree nodes is too weak for having the pure power-law. Instead degree distribution converge to stretched exponential.

## Ageing model

In real systems preferential attachment can be more influenced by the age of the node. If parameter alpha is negative, ageing effect overcomes the role of preferential attachment, and scale-free properties are lost. For large negative alpha, the network turns into the chain, where the youngest nodes are the most attractive. On the other hand for a positive alpha, new nodes will link to older nodes. Positive alpha makes the network more heterogeneous, and scale-free nature still exist but exponent gamma is different from 3. for the high alpha all nodes will tend to connect to oldest node.

The model starts with small number of nodes randomly connected. Further, at each time step new node arrives in the network and makes connection with one old node, already present in the network. The way in which new nodes are linked is governed by various mechanisms. They can have preference to nodes with high degree (preferential attachment), or preference to nodes with specific age. In the network model with aging nodes the probability that link is created between two nodes depends on the node degree and age [?]:

$$\Pi_i(t) \sim k_i(t)^\beta \tau_i^\alpha \quad (1.19)$$

where  $k_i(t)$  is a degree of a node  $i$  at time  $t$ , and  $\tau_i$  is age difference between node  $i$  and newly added node.

The values of model parameters  $\beta$  and  $\alpha$  control the topology of networks. For example if we fix  $\beta = 1$  and  $\alpha = 0$  generated networks are scale-free; degree distribution is  $P(k) \sim k^{-\gamma}$  with  $\gamma = 3$ , while in the case of nonlinear preferential attachment  $\beta \neq 1$  and  $\alpha = 0$  scale-free properties disappear. Scale-free property can be produced along the critical line  $\beta(\alpha^*)$  in the  $\alpha - \beta$  phase diagram, see Figure 1.10. For  $\alpha > \alpha^*$  networks have gel-like small world behavior, while for  $\alpha < \alpha^*$  but close to line  $\beta(\alpha^*)$  networks have stretched exponential shape of degree distribution [?].

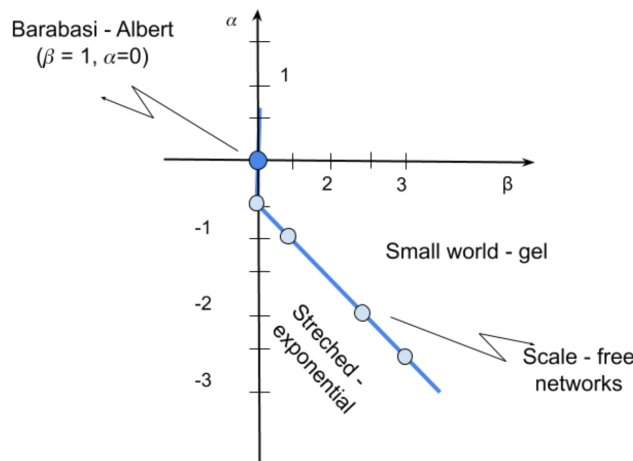


Figure 1.10: Phase diagram of aging network model

The networks generated with constant growth signal are uncorrelated trees.

In the general ageing model, we have linking rules where rules connecting probability depends on both of node degree and age difference between new and old node. With parameters alpha and beta we can control the structure of generated networks. I already talked about some limits of the general model. We saw that for specific parameters there are SF networks, BA model, if we move from that point SF behaviour with power-law with exponent 3 is lost. And other classes of networks can appear. In general, model, when alpha and beta are both positive, rich get richer phenomena is more promoted. On the other hand, the region where beta is positive and alpha negative can be interesting, because SF networks can appear only along the critical line.

In growing network models is considered that at each time step one node is added to the network. The remaining question is if there is any change if network growth is not linear anymore and how does it influence the structure of obtained networks. In this work, we use numerical simulations to explore the case when  $M(t)$  is a correlated time-varying function and study how these properties influence the structure of generated networks for different values of parameter  $-\infty < \alpha \leq -1$  and  $\beta \geq 1$  and constant  $L$ .