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.

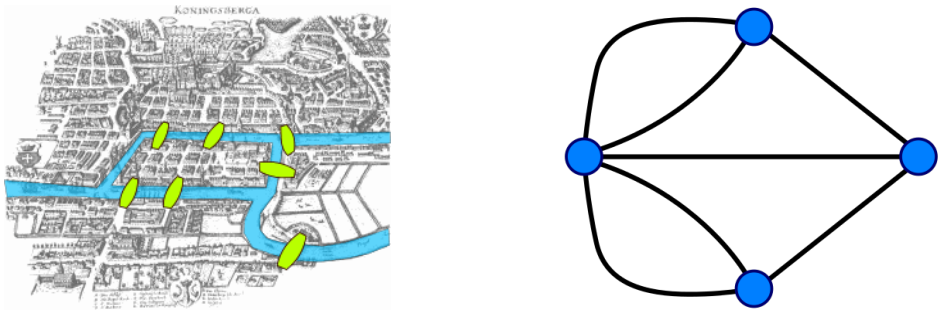

Figure 1.1: The Kronigsber problem of seven bridges.

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 $i, j \in E$. 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:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix} \tag{1.1}$$

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.

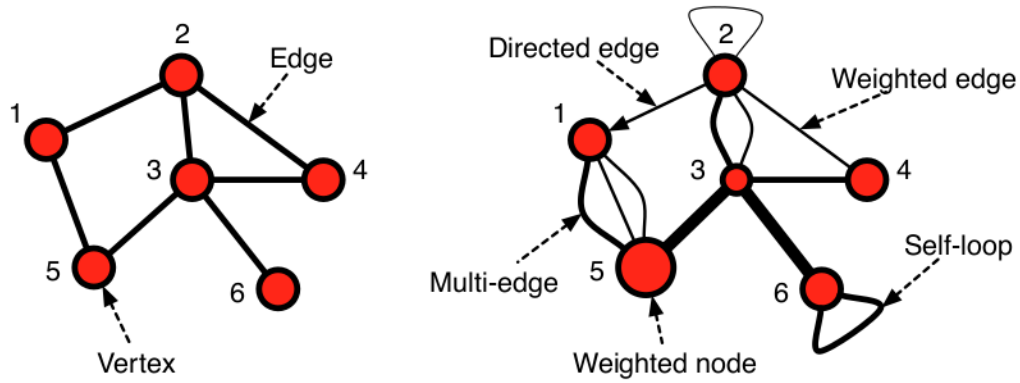

Figure 1.2: Graph types

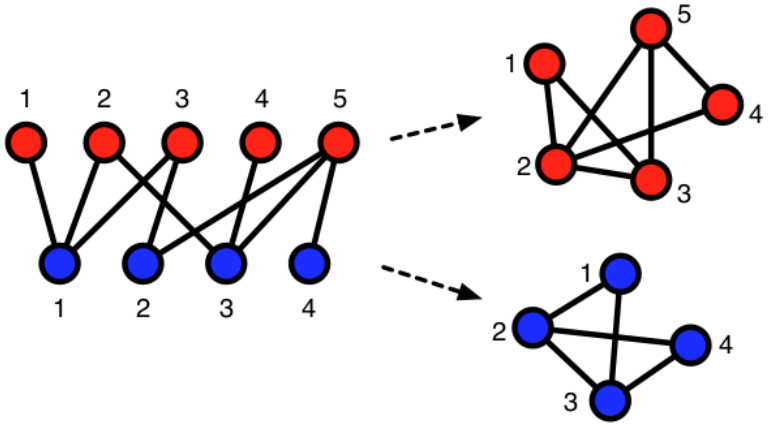

Figure 1.3: Graph types

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 gxn, having elements B_{ij} 1 if persoon 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.

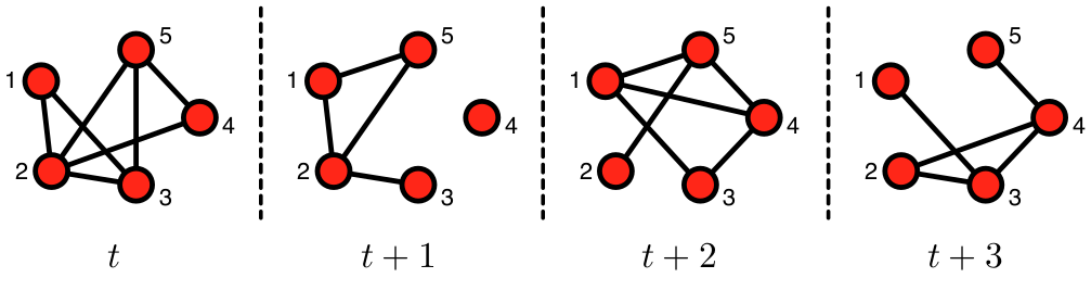

Figure 1.4: Graph types

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 Δt 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 $G = G(1), ...G(t_{max})$, 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.

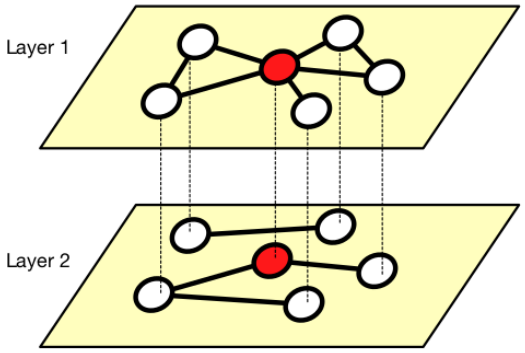

Figure 1.5: Graph types

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 × V × 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, $i_1 i_2, .., i_k$ 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 id 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 where g ij = 1 for all i 6 = 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 di¤erent 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. 39 This follows since each time a node is "entered" by one link on the walk it must be "exited"by a di¤erent 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 su¢ cient for there to exist such a closed walk.

A connected network g has a closed walk that involves each link ex- actly 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 …nd 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 chal- lenging 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 γ 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 γ 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 kth 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.

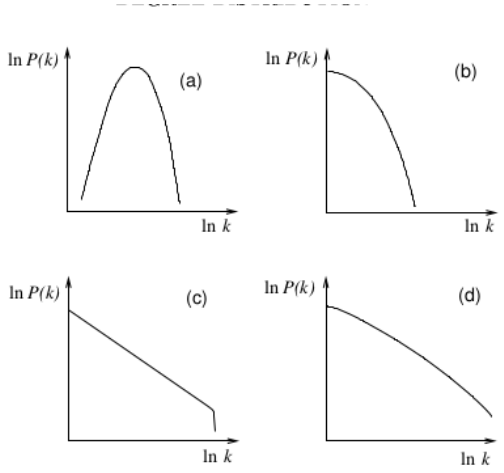

Figure 1.6: Distributions, ovde hocemo 3 distribucije+ linearna vs. logarithm scala, takodje pokazati kako izgleda frequency order

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 $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$. The correlation coefficient $r(x, y)$ between x and y is defined as:

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

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 $V(G) = \{v_1, ..v_n\}$, $\boldsymbol{A}[i,j] = 1$ if there is a link between nodes v_i and v_j. If G is a simple graph with adjacency matrix $\boldsymbol{A}$ and degree sequence $\boldsymbol{d} = [d_1, ..., d_n]$

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

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; dissasortative 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:

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

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

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

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

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

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 **[?]**

$$D(G, G^{'}) = \omega \left| \sqrt{\frac{J(P_1, ..P_N)}{log(d)}} - \sqrt{\frac{J(P_1^{'}, ..P_N^{'})}{log(d^{'})}} \right| + (1 - \omega) \sqrt{\frac{J(\mu_G, \mu_{G^{'}})}{log 2}} \tag{1.7}$$

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

$$J(P_1, .., P_N)) = \sum_{i,j} p_i(j) log(\frac{p_i(j)}{\mu_j}) \tag{1.8}$$

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 $J(P_1, ..., P_N)$ 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,

$$c_i = 2e_i/(k_i(k_i - 1)) \tag{1.9}$$

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 $l << N$, 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

Barabasi-Albert model

Nonlinear BA model

Aging model