

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. [1]

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.

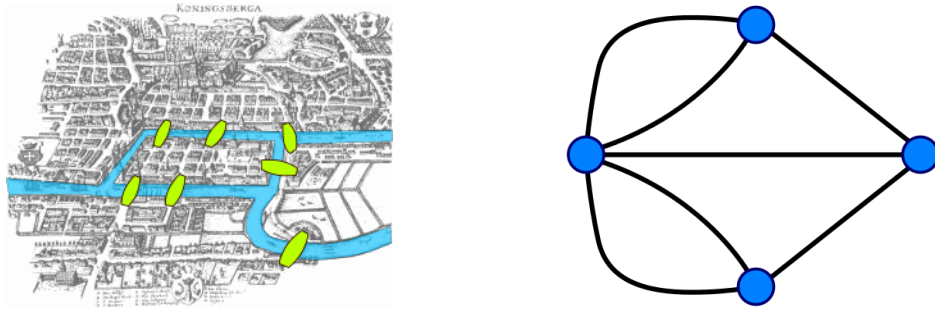


Figure 1.1: The Kronigsber problem of seven bridges.

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 [2]

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.

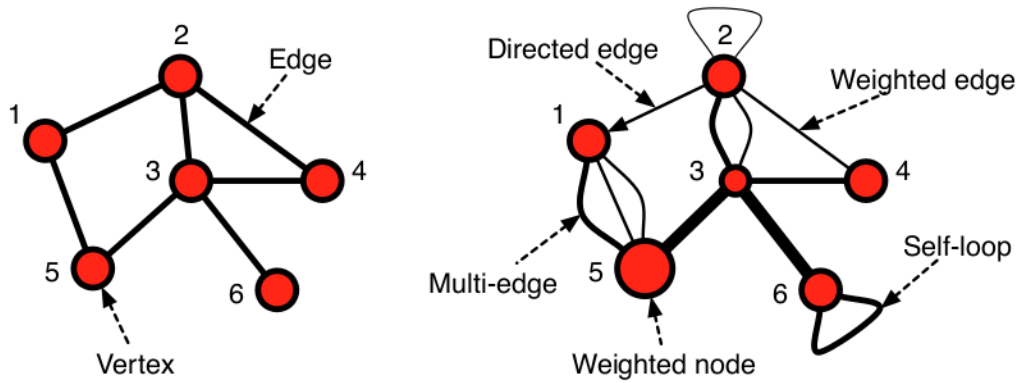


Figure 1.2: Graph types

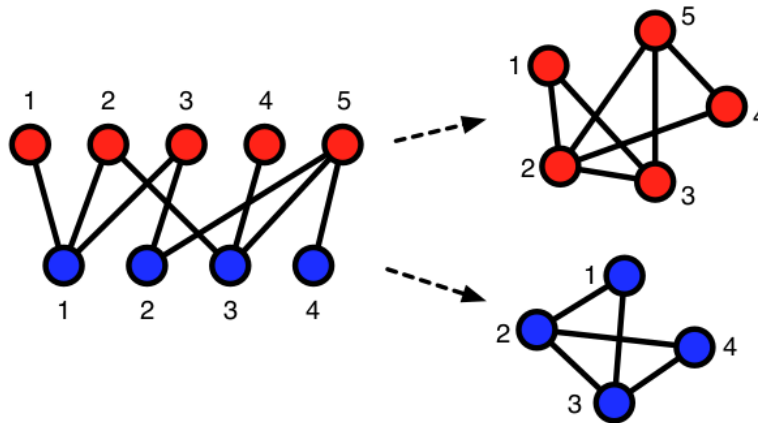


Figure 1.3: Graph types

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 that any projected network P has at least one bipartite network such that the projection of B results in P .

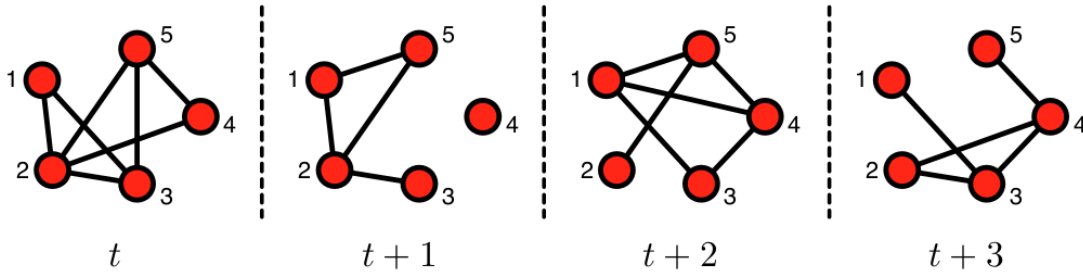


Figure 1.4: Graph types

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 $G = G(1), \dots, G(t_{max})$, 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.

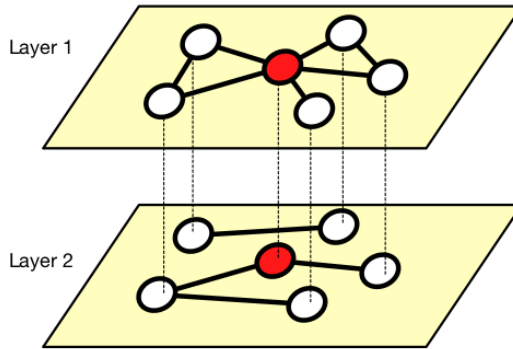


Figure 1.5: Graph types

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

Thus the ability to find groups or clusters in a network can be a useful tool for revealing structure and organization within networks at a scale larger than that of a single node or a few nodes. The occurrence of groups or communities is not limited to social networks. Clusters of nodes in a web network, for instance, might indicate groups of related web pages. Clusters of nodes in a metabolic network might indicate functional units within the network. The ability to find groups also has another practical application: it allows us to take a large network and break it apart into smaller subsets that can be studied separately. The network in Fig. 14.1 is quite small, but others may be much larger, millions of nodes or more, making their analysis and interpretation challenging. Breaking such networks into their component clusters is a useful technique for reducing them to a manageable size. One example of this approach is in network visualization. A network with a million or more nodes can rarely be visualized in its entirety, even with the aid of good visualization software. Such networks are simply too big to be represented usefully on the screen or on paper. If the nodes of the network divide naturally into groups, however, then we can make a simpler but still useful picture by representing each group as a single node and the connections between groups as edges. An example is shown in Fig. 14.2. This simplified representation allows us to see the large-scale structure of the network without getting bogged down in the details of the individual nodes. If one wanted to see the individual nodes, one could then “zoom in” to a single group and look at its internal makeup. The problem of finding groups of nodes in networks is called community detection. Simple though it is to describe, community detection turns out to be a challenging task, but a number of methods have been developed that return good results in practical situations.

Core-periphery structure

Core-periphery structure describes a network whose nodes are divided into two community, densely connected core and less connected periphery. If we consider the average probabilities of edges within each group as p_{11} and p_{22} , and between groups p_{12} , instead of traditionally assortative or disassortative structure we can define core-periphery structure $p_{11} > p_{12} > p_{22}$. In the principle core-periphery structure does not have to be limited to only two groups, and we can define layered, onion, structure. The network can have more cores, that are not directly connected to each other.

The simple method for finding core-periphery structure is to assume that nodes in core have higher degree in the core than in the periphery. Another simple method is to construct k-cores. K core is group of nodes that each has connection to at least k other members of the group. K-cores form a nested set, and become denser with higher k. The core-periphery structure can be detected optimizing the measure similar to modularity, as defined by Borgatti and Everett. Their goal is to find the division that minimizes the number of edges in the periphery. So they define the score function that is equal to number of edges in the periphery minus the expected number of such edges placed at random. $\rho = \frac{1}{2} \sum_{ij} (A_{ij} - p) g_i g_j$. They used genetic algorithm to minimize this function.

The another way to detect core-periphery structure is to use the inference method based on fits to a stochastic block model. In this method we fit observed network to a block model with two groups, such that edge-probabilities have form $p_{11} > p_{12} > p_{22}$. The only downside of this model is that method is going to find the structure that optimize likelihood, and we can not say weather it is core-periphery or community structure.

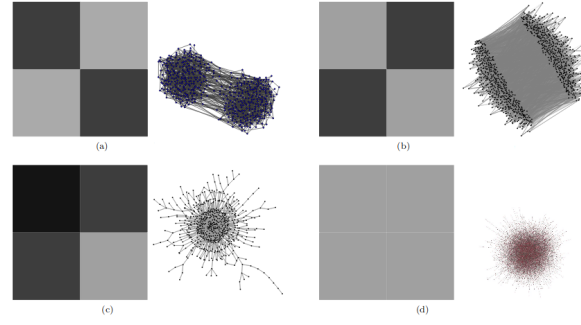


Figure 1.6: Stochastic Block model for different networks structures. (a) assortative. (b) dissortative. (c) core-periphery. (d) Erdos Renyi random graph.

Stochastic block model

The network or graph is the structure of nodes and edges, where each edge connects two nodes. Nodes can be organized into groups, called communities. Identifying these hidden blocks can lead to interesting insights into the network. However, the community detection problem does not give a precise definition of what a community is. As a consequence, many approaches try to recover such structural patterns in the network [4].

A common definition of a community is that it is densely connected subgraph [5]. We can find these subgraphs by optimizing an objective function, such as modularity function. It measures the difference in the number of edges between the given network and the network with the same number of nodes but randomly connected. In this approach, we try to maximize the density of connections inside a group by focusing more on assortative¹ group structures.

Another type of networks is the bipartite network that has two disjoint sets of nodes. The edges exist only between nodes from different sets. Networks of this class can appear in real-world data, such as users-movies preference, collaboration network for scientists and papers, etc. Application of density-based approach requires to first project bipartite network to one of its partitions and then find communities in that projection. With this, some information is lost. On the other hand, the method that is directly applicable to bipartite networks is Stochastic Block model, from which the models considered in this paper are derived.

Stochastic block model (SBM) is based on connection probabilities between nodes. It is a generative model which includes existence of communities. Parameters that describe SBM for network G with N nodes are:

- k : number of groups
- group assignment vector, g : $g_i \in \{1, 2, \dots, k\}$, gives the group index of node i .
- SBM matrix, $p_{k \times k}$, whose elements p_{ij} are the probabilities that edges between groups g_i and g_j exist.

Note that nodes within one group have the same connection probabilities.

SBM can generate and describe different types of network structures. Figure 1.6 [5] shows how the model matrix corresponds to resulting networks with two communities. First, for

¹Networks where nodes tend to connect with other nodes of a similar degree. Edges are more likely inside blocks than out of them.

the assortative network (1.6 a), diagonal elements of the matrix have higher probabilities. This indicates dense connections inside the group, just like in classic community structures. In disassortative structure, (1.6 b), more connections exist between two partitions than inside them, i.e. off-diagonal elements have higher probabilities. Bipartite networks can be represented like this.

Figure (1.6 c) shows how the model represents core-periphery networks. Nodes of one block (core) are well connected with itself and with other partition (periphery). From the last case, we can note that SBM with one group is the Erdos Renyi random graph (1.6 d) because all probabilities inside and between groups are equal.

The benefit of this model is that we can generate many networks with similar group structure. The model can fit real data, which results in finding network communities. For the given network G and number of groups k , the best nodes partition g is found by maximizing the likelihood function. Beside inferring communities, SBM has application in prediction of missing links. This simply formulated model has many variants, motivated by specific properties of real data. For example, for networks which are degree heterogeneous, there is degree corrected SBM. In some social networks, users can belong to more than one group, and this can be modelled with mixed membership SBM. Other extensions include application to bipartite, weighted network, hierarchical model, etc. Also, several algorithms for optimization of likelihood function are proposed. The overview of these versions and methods are given in [6]. In this paper, we will focus on Single and Mixed Membership SBM applied on bipartite networks.

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.

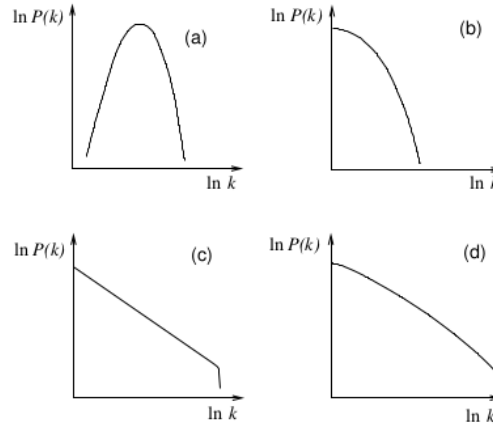


Figure 1.7: Distributions, ovde hocemo 3 distribucije+ linearna vs. logarithm skala, takodje pokazati kako izgleda frequency order

- 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 [3]. 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.

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 [2], $kP(k'|k)P(k) = k'P(k|k')P(k')$ [2]. 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 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 .

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

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 [16]

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

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

A common way toward spreading information is simply having a node up- date its neighbors. In turn, neighbors can inform their neighbors, and so on. There are many variations to this model, such as having a node select only one or a few of its neighbors, or deciding to stop spreading updates when it notices that a selected neighbor already has the information. Informally, this type of dissemination is often described in the form of gossiping models, also known as epidemic dissemination [Eugster et al., 2004]. The model is very general: instead of information we can also consider spreading of dis- eases, but also viruses over the Internet. Another example is that of forming of opinions, which often depends on what the majority of your community thinks. We shall return to these issues in more detail when discussing peer- to-peer networks in Chapter 8. When considering real-world networks, we often see that they are orga- nized as a collection of interconnected groups. In terms of social networks, this means that we can often clearly distinguish communities of nodes with many links between its members, yet relatively few links between nodes that belong to different communities. Actually indicating which nodes be- long to which communities may not be easy at all. Also, nodes generally belong to more than one community. However, we can express the existence of communities by means of a clustering coefficient. As shown by Xu and Liu [2008], it turns out that there is a clear relationship between the speed by which information is disseminated in social networks and the clustering co- efficient: the higher the degree of clustering, the slower the dissemination. To a certain extent, this result may seem quite obvious, but from a formal (i.e., mathematical) point of view, it turns out to be not so trivial. What this means is that if we want to design a dissemination protocol, we may need to take special measures in highly clustered networks in order to guarantee a certain performance regarding the dissemination speed. This alone has been enough reason for researchers to define and measure the clustering coefficient of a network. Besides this reason, measuring the clustering coefficient obviously allows us to simply compare different

networks, without necessarily wanting to make use of the actual values of the respective coefficients. In this sense, clustering coefficients can help in classifying networks.

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.

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.

Bibliography

- [1] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.
- [2] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.
- [3] M. E. J. Newman. *Networks: An Introductio*. Oxford University Press, 2010.
- [4] Martin Rosvall, Jean-Charles Delvenne, Michael T. Schaub, and Renaud Lambiotte. Different approaches to community detection. *CoRR*, abs/1712.06468, 2017.
- [5] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1 – 44, 2016. Community detection in networks: A user guide.
- [6] Thorben Funke and Till Becker. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE*, 14(4):1–40, 04 2019.
- [7] Jan W Kantelhardt, Stephan A Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):87–114, 2002.
- [8] Espen Alexander Fürst EAFI Ihlen. Introduction to multifractal detrended fluctuation analysis in matlab. *Frontiers in physiology*, 3:141, 2012.
- [9] Claudiu Duma, Nahid Shahmehri, and Germano Caronni. Dynamic trust metrics for peer-to-peer systems. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, pages 776–781. IEEE, 2005.
- [10] A. Melnikov, J. Lee, V. Rivera, M. Mazzara, and L. Longo. Towards dynamic interaction-based reputation models. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 422–428, 2018.
- [11] Hernán A Makse, Shlomo Havlin, Moshe Schwartz, and H Eugene Stanley. Method for generating long-range correlations for large systems. *Physical Review E*, 53(5):5445, 1996.
- [12] Jan W Kantelhardt, Eva Koscielny-Bunde, Henio HA Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3-4):441–454, 2001.
- [13] C-K Peng, Sergey V Buldyrev, Shlomo Havlin, Michael Simons, H Eugene Stanley, and Ary L Goldberger. Mosaic organization of dna nucleotides. *Physical review e*, 49(2):1685, 1994.
- [14] Kamalika Basu Hajra and Parongama Sen. Phase transitions in an aging network. *Physical Review E*, 70(5):056103, 2004.

- [15] Sergey N Dorogovtsev and José FF Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63(5):056125, 2001.
- [16] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M Pardalos, Cristina Masoller, and Martín G Ravetti. Quantification of network structural dissimilarities. *Nature communications*, 8(1):1–10, 2017.
- [17] Marija Mitrović and Bosiljka Tadić. Emergence and structure of cybercommunities. In *Springer Optimization and Its Applications*, volume 57, pages 209–227. Springer International Publishing, 2012.
- [18] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.
- [19] Milovan Suvakov, Marija Mitrovic, Vladimir Gligorijevic, and Bosiljka Tadic. How the online social networks are used: dialogues-based structure of myspace. *Journal of The Royal Society Interface*, 10(79):20120819, 2013.
- [20] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.
- [21] Marc Barthélemy. *The structure and dynamics of cities*. Cambridge University Press, 2016.
- [22] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.
- [23] Jelena Smiljanić, Arnab Chatterjee, Tomi Kauppinen, and Marija Mitrović Dankulov. A theoretical model for the associative nature of conference participation. *PloS one*, 11(2):e0148528, 2016.
- [24] Ali Montazeri, Soghra Jarvandi, Shahpar Haghighat, Mariam Vahdani, Akram Sajadian, Mandana Ebrahimi, and Mehregan Haji-Mahmoodi. Anxiety and depression in breast cancer patients before and after participation in a cancer support group. *Patient education and counseling*, 45(3):195–198, 2001.
- [25] Kathryn P Davison, James W Pennebaker, and Sally S Dickerson. Who talks? the social psychology of illness support groups. *American Psychologist*, 55(2):205, 2000.
- [26] Wendy K Tam Cho, James G Gimpel, Daron R Shaw, et al. The tea party movement and the geography of collective action. *Quarterly Journal of Political Science*, 7(2):105–133, 2012.
- [27] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.
- [28] Sandra González-Bailón, Javier Borge-Holthoefer, and Yamir Moreno. Broadcasters and hidden influentials in online protest diffusion. *American behavioral scientist*, 57(7):943–965, 2013.
- [29] János Török, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo Kaski, and János Kertész. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters*, 110(8):088701, 2013.
- [30] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PloS one*, 7(6):e38869, 2012.

- [31] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.
- [32] Arnab Chatterjee, Marija Mitrović, and Santo Fortunato. Universality in voting behavior: an empirical analysis. *Scientific reports*, 3(1):1–9, 2013.
- [33] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.
- [34] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.
- [35] Jelena Smiljanić and Marija Mitrović Dankulov. Associative nature of event participation dynamics: A network theory approach. *PloS one*, 12(2):e0171565, 2017.
- [36] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.
- [37] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016, 2009.
- [38] Luís A Nunes Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, H Eugene Stanley, and Michael HR Stanley. Scaling behavior in economics: I. empirical results for company growth. *Journal de Physique I*, 7(4):621–633, 1997.
- [39] Michael HR Stanley, Luis AN Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, and H Eugene Stanley. Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806, 1996.
- [40] Giorgio Fazio and Marco Modica. Pareto or log-normal? best fit and truncation in the distribution of all cities. *Journal of Regional Science*, 55(5):736–756, 2015.
- [41] Konglin Zhu, Wenzhong Li, Xiaoming Fu, and Jan Nagler. How do online social networks grow? *Plos one*, 9(6):e100023, 2014.
- [42] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 673–682, 2012.
- [43] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE*, 9(1):1–11, 01 2014.
- [44] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.
- [45] Hernan Mondani, Petter Holme, and Fredrik Liljeros. Fat-tailed fluctuations in the size of organizations: the role of social influence. *PLoS One*, 9(7):e100527, 2014.

- [46] Dongfeng Fu, Fabio Pammolli, Sergey V Buldyrev, Massimo Riccaboni, Kaushik Matia, Kazuko Yamasaki, and H Eugene Stanley. The growth of business firms: Theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences*, 102(52):18801–18806, 2005.
- [47] Gerald F Frasco, Jie Sun, Hernán D Rozenfeld, and Daniel Ben-Avraham. Spatially distributed social complex networks. *Physical Review X*, 4(1):011008, 2014.
- [48] Jiang-Hai Qian, Qu Chen, Ding-Ding Han, Yu-Gang Ma, and Wen-Qing Shen. Origin of gibrat law in internet: Asymmetric distribution of the correlation. *Physical Review E*, 89(6):062808, 2014.
- [49] Marija Mitrović, Georgios Paltoglou, and Bosiljka Tadić. Quantitative analysis of bloggers’ collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02005, 2011.
- [50] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.
- [51] Ana Vranić and Marija Mitrović Dankulov. Growth signals determine the topology of evolving networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(1):013405, 2021.
- [52] Akрати Saxena and Harita Reddy. Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *Journal of Computational Social Science*, pages 1–33, 2021.
- [53] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Activity archetypes in question-and-answer (q&a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing*, 2(1):1–23, 2019.
- [54] Rogier Slag, Mike de Waard, and Alberto Bacchelli. One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 458–461. IEEE, 2015.
- [55] Anamika Chhabra and S RS Iyengar. Activity-selection behavior of users in stackexchange websites. In *Companion Proceedings of the Web Conference 2020*, pages 105–106, 2020.
- [56] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. The size conundrum: Why online knowledge markets can fail at scale. In *Proceedings of the 2018 World Wide Web Conference*, pages 65–75, 2018.
- [57] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Self-and cross-excitation in stack exchange question & answer communities. In *The World Wide Web Conference*, pages 1634–1645, 2019.
- [58] Yaniv Dover, Jacob Goldenberg, and Daniel Shapira. Sustainable online communities exhibit distinct hierarchical structures across scales of size. *Proceedings of the Royal Society A*, 476(2239):20190730, 2020.
- [59] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczkai, Marián Boguná, Guido Caldarelli, et al. Quantifying randomness in real networks. *Nature communications*, 6(1):1–10, 2015.

-
- [60] Damon Centola, Víctor M Eguíluz, and Michael W Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.
- [61] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
- [62] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *arXiv preprint arXiv:2005.10191*, 2020.
- [63] Ekaterina Yashkina, Arseny Pinigin, JooYoung Lee, Manuel Mazzara, Akinlolu Solomon Adekotujo, Adam Zubair, and Luca Longo. Expressing trust with temporal frequency of user interaction in online communities. *Advances in Intelligent Systems and Computing*, pages 1133–1146, Cham, 2020. Springer International Publishing.
- [64] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):1–16, 2012.
- [65] Naomi A Arnold, Benjamin Steer, Imane Hafnaoui, Hugo A Parada G, Raul J Mondragon, Félix Cuadrado, and Richard G Clegg. Moving with the times: Investigating the alt-right network gab with temporal interaction graphs. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–17, 2021.
- [66] Vincent Labatut and Hocine Cherifi. Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*, 2012.
- [67] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, 2005(09):P09008, 2005.
- [68] Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer, 2009.
- [69] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.