Chapter 1

Methodology

1.1 Complex networks

Many real systems are composed of a large number of elements interacting with each other. Due to interactions, without any central force, the system exhibits the emergence of collective behaviour on the macro level. Such a system is called a Complex System and its properties can not be predicted from the behaviour of the one individual. An example of a complex system is the human brain. The structure of the brain network and its properties are fundamental for brain functioning, while an emergent phenomenon is a human intelligence. In societies, people's interactions lead to civilization, economy, formation of social groups. Also, the animal populations show different levels of organization that emerge from the individual's interactions [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. For the brain network, we can construct representation with neurons and synapses, representing the brain connectivity. Neurons in the same brain area are closely connected [2]. Similarly, we can define 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 (vertices) and links (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) [3, 4]. 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. 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, as in figure 1.1, 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. By this it is possible to

enter one part of the land from one bridge and leave it by the other. As each node has odd number of connections, in this case it is not possible, see Fig. 1.1.

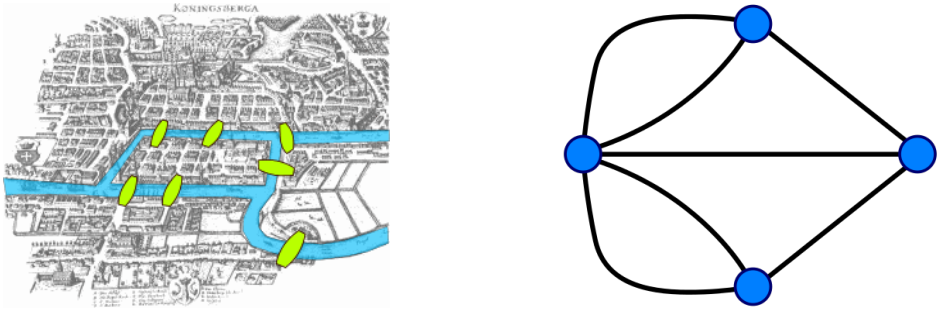

Figure 1.1: The Kronigsber problem of seven bridges.

1.2 Types of networks

The graph or network G is defined as $G = (\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V} = \{v_1, v_2, ...v_N\}$ is a set of N nodes (vertices), and $\boldsymbol{E} = \{e_1, ..e_L\}$ is a set of L edges (links). The edge is pair of nodes $e = (v_i, v_j)$, such that $\{v_i, v_j\} \in \boldsymbol{V}$. The most basic network representation considers **unweighted and undirected** structure. The edges are unweighted, meaning that all interactions in the network are equally important. Because network is un-directed, edges are symmetric, such that (v_i, v_j) implies (v_j, v_i). In **directed** networks this simmetry is broken. The interaction between two nodes v_i and v_j, can be only in one direction. A typical example is World Wide Web, where webpages are nodes and hyperlinks are directed edges. In biological networks, gene regulation and neural activation can be described as directed network. The first column a) in Figure 1.2 shows the graphical representation of two networks with equal number of nodes; the first one is underected and the second one is directed.

Even though, graphical representation can be useful for describing the network structure, mathematical representation allow us to characterize the statistical properties of the networks. The graph G, with N nodes could be represented with **adjacency matrix** $|A| = N \times N$ [1]. The elements of the matrix are positive if there is connection between two nodes v_i and v_j.

$$A_{ij} = \begin{cases} 1 & (v_i, v_j) \in E \\ 0 & (v_i, v_j) \notin E \end{cases} \tag{1.1}$$

The column b) on Figure 1.2 shows adjacancy matrix representation of given graphs. By convention diagonal elements $A_{ii} = 0$, as self-loops are not allowed. For undirected network adjacency matrix is symmetric $A_{i,j} = A_{ji}$, but in the case of directed network matrix is not symmetric, as edges are drawn in one direction only.

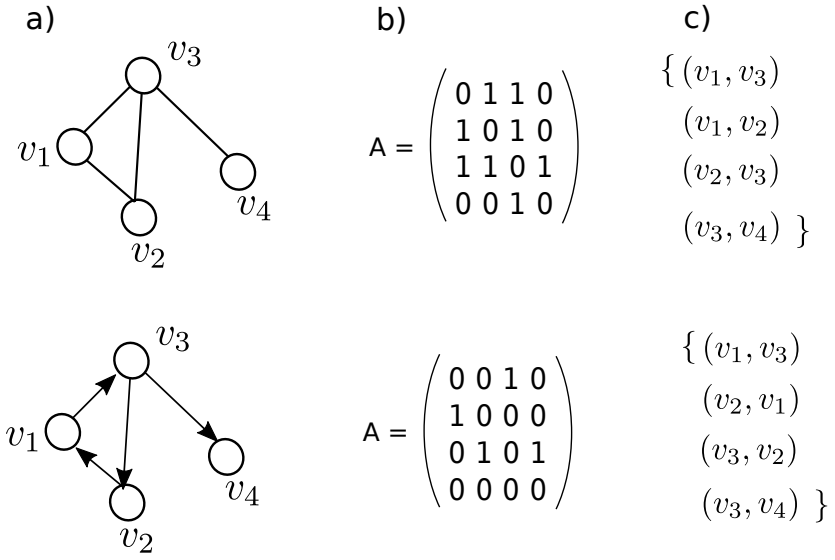

Figure 1.2: a) Graph representation of undirected (top panel) and directed (bottom panel) network. The same networks are represented with adjacency matrices column b), and edge list representation in column c).

The number of edges and nodes are dependent variables. Considering that each node can make $N-1$ connections, the maximum number of the edges in the network is $L_{max} = N(N-1)/2$, as each edge is counted twice. For directed network it is possible to draw $L_{max} = N(N-1)$ edges [5]. When it comes to large networks, they are sparse, meaning that the number of links is $L << L_{max}$. As consequence, the adjacency matrix is also sparse structure (has many zeros) that takes large portion of computer memory [6]. It is common to represent the graph as edge list. In this case, illustrated on Figure 1.2, column c), graph is described with the list of links that are in the graph, $G = \{\{v_i, v_j\}\}$. Still with this representation we are not able to distinguish between directed and undirected graph structures, so in the computational algorithm should be specified if the edges are considered symetric or not.

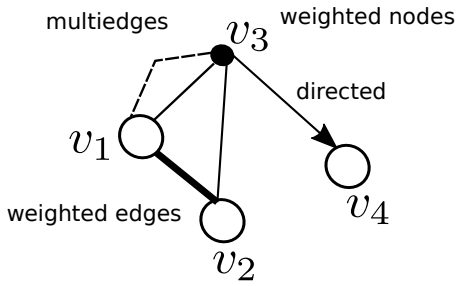

Figure 1.3: The complex networks may represent different characteristics of the system. The edges can be directed, weighted or multiply. Also nodes can be assigned with different weights or any relevant feature.

To create the more realistic models, sometimes is essential to include the specific properties of the system in the network representation. For example, to emphasis the frequent interactions between

nodes, edges can be assigned with different values, such networks are **weighted**. They can be described with adjacency matrix, whose elements can take any real number $A_{ij} = w_{ij}$ and $w_{ij} > 0$. In general edges may be associated with any categorical variable. Similarly additional properties can be added to nodes, or even to the whole network structure. To include the **temporal** component in the network, edges are characterized with the time when the interaction between nodes happen. Finally, if two nodes interact in different ways, the **multigraph** is appropriate configuration where multiply edges are allowed. The graphical representation of discussed network representations is given on the Figure 1.3.

A **bipartite network** consists of two types nodes. The nodes in the same partition are not connected, while links exist only between partitions. For many real systems, a bipartite graph is a natural representation[6, 2]. For example, the bipartite network of people and groups has two distinct node partitions while links indicate the memberships. Another example is a system of customers and products. The link between user and item is created when the user buys an item. The bipartite networks find their application in the algorithms for recommended systems, whose goal is to recommend items that may interest the user. Actually, to find the most probable missing links in the network.

In a bipartite network, nodes in one partition are not connected. Still, we can analyse a single node type if we project the bipartite network on one partition. The primary assumption is that two nodes in one partition could be connected if they point to the same node in another partition. Consider the network of movies and actors. The one mode projection of movies is an undirected network whose links indicate that two movies share the same actors. On the other hand, another projection is a network of actors. The links exist if two actors appear in the same movie [7, 6].

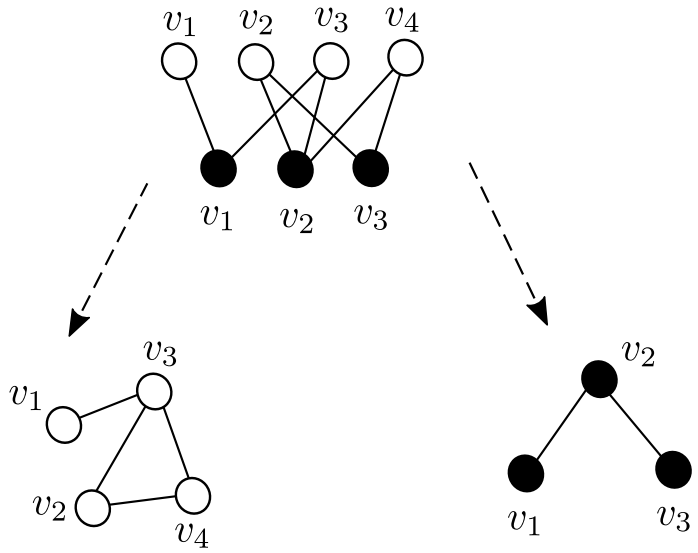

Figure 1.4: Bipartite graph and two partition projections.

We should be aware that important information is lost when creating a one-mode projection. First of all, without having weighted edges in the network of actors, it is impossible to have information on how many movies two actors appear in. From the one-mode projection, we can not reconstruct the original network. Moreover, two different bipartite networks may have the same projected networks. The important consequence of the network projection is the creation of cliques; subgraphs where all nodes are connected.
In general, it is possible to define the k-bipartite network. The same rules apply as before. There are k distinct node partitions, while the edges exist only between different types of nodes.

Temporal networks. Studying the real systems as static networks can give us a lot of insight into the system properties. Still, real systems are not static; they evolve not only in the number of elements but also in the number of interactions between them. Some interactions in the system may repeat in different intervals and could be described with complex activity patterns. Including time dimension in the network representation allows us to study the properties of the system closely. The temporal information may matter a lot [8]. For example if interaction between nodes (v_1, v_2) happened before in time than (v_2, v_3), then nodes v_1, v_3 would not be connected, as it is the case in the static network.

The temporal network is a collection of timestamped edges. Each edge is defined as $(v_i, v_j, t, \Delta t)$, where v_i and v_j, are nodes t is time when interaction happen, and Δt is event duration [9]. The duration of the events may vary, as in the phone-call network. Also, for many systems, the time resolution of event duration is too small. For example, this parameter may be neglected when people interact on social platforms or email each other because the event time is too short, it scales in seconds.

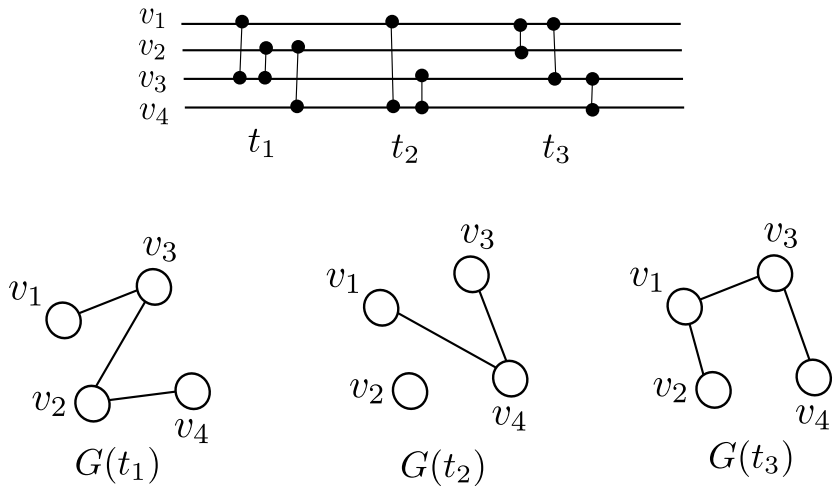

Figure 1.5: Temporal network.

The temporal network can be represented as sequence of static networks that evolve in time, $G = \{G(t_1), G(t_2), ..., G(t_{max})\}$. At each time step, we can create the network and analyze the macroscopic properties of the given network snapshot. With this, we can end up with graph snapshots with many disconnected components or empty graphs for some points [10]. Sometimes, a much better approach is to aggregate the links that over time-windows. Here, we need to specify the time window length w. Interactions in the time interval $0 \leq t < w$ enter the first snapshot. The next snapshot takes edges $w \leq t < 2w$, and so on. The time windows are not overlapping, but generally, it is possible to slide the time window for different periods $1 \leq \delta t \geq w$. The downside of this method is that we can not recover original data points. The larger the time window is, the more information is lost. If the time window is set to $w = t_{max}$, there is only one snapshot, and the temporal data are no more available [11, 12].

Multilayer networks were introduced for studying systems in which different types of interaction exist. This formalism allows one to investigate diverse network systems and to combine different types of data into one model [13]. In a multilayer or multiplex network, all nodes are present in each layer, but their interactions among layers differ. Two nodes may be connected in one layer but not in the other. Different online social systems may be an example of a multiplex network when users are connected on one platform but not on the other [14]. Or the airline transportation network, where each layer represents the flights of different airline companies [15].

1.3 The structure of complex networks

1.3.1 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 same, then graphs 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. Here we summarize the forms of degree distributions that are mostly found in the complex network theory:

- 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 many 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 $\gamma > 2$. Therefore, real networks have a scale-free structure with the emergence of the hubs [7].

When plotting the degree distribution, it is common to use scaling of the axis. As many nodes have low degree, like for power-law or exponential distribution it is more useful to use logarithmic scale. Now it is more easily notices that data-points follow straight line, meaning that degree distribution is some kind of exponential function.

1.3.2 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 does 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 [1], $kP(k'|k)P(k) = k'P(k|k')P(k')$ [1]. 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.

1.3.3 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,

$$c_i = 2e_i/(k_i(k_i - 1)) \tag{1.5}$$

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.

1.3.4 Network paths

In the network structure, the interacting nodes are directly connected with the edge. In this representation we can say that distance between them is $d_{v_i,v_j} = 1$. Distance defined like this does not have any physical meaning. Its purpose is to describe how the position of nodes in the network structure influences the other distant nodes.

The **path** between two nodes, v_i and v_j is a sequence of edges $\{(v_1, v_2), (v_2, v_3), ...(v_k, v_{k+1})$ $, ...(v_{n-1}, v_n)\}$, where $v_1 = v_i$, $v_n = v_j$. In the path, the nodes are distinct. Otherwise, the sequence is called a **walk**, where each node can be visited many times. Also, it is possible to define a **cycle**, a path that starts and ends on the same node while other nodes in the cycle are distinct. The length of the path, walk or cycle is the number of links in the sequence. Using the adjacency matrix we can easily calculate the number of walks between two nodes. The A^2 gives us walks of length 2, the A^3, number of walks of length 3, and so on.

The network is connected if it is possible to define the path between every two nodes in the network. When it is not the case, the network is disconnected into two or more connected components. Note that the component can be an isolated node. Also, in directed networks may happen that node v_i is reachable from node v_j, but if we start from v_j we can not find the path to the v_i. Such a graph is connected but is called a weakly connected component.

We can find different paths between two nodes in the network, but the most important one is the **shortest path**. The distance between two nodes $d(v_i, v_j)$ is defined as the shortest path length between two nodes. In the case of weighted networks, it is the path with minimal weight, and the length of such path does not have to be minimal. Distances on the network can give us insight into how similar networks are and indicate the node's relative importance in the network.

The **radius** is the minimum overall eccentricity values, while the **diameter** defines the largest distance between nodes in the network. These definitions apply to directed and undirected graphs.

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

$$\bar{d}(u) = \frac{1}{|V| - 1} \sum_{v \in V, v \notin u} d(u, v) \tag{1.6}$$

From there, the **average path length** is mean value over $\bar{d}(u)$.

$$\bar{d}(G) = \frac{1}{|V|} \sum_{u \in V} \bar{d}(u) \tag{1.7}$$

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

1.3.5 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]

$$D(G, G^{'}) = \omega \left| \sqrt{\frac{J(P_1, ..P_N)}{log(d)}} - \sqrt{\frac{J(P_1^{'}, ..P_N^{'})}{log(d^{'})}} \right| + (1 - \omega) \sqrt{\frac{J(\mu_G, \mu_{G^{'}})}{log2}} \tag{1.8}$$

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

$$J(P_1, .., P_N)) = \sum_{i,j} p_i(j) log(\frac{p_i(j)}{\mu_j}) \tag{1.9}$$

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

The first term in equation 1.8 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.

1.3.6 Community structure

Nodes can be organized into groups, called communities. Identifying these hidden blocks can lead to interesting insights into the network. The communities are expected in social networks, as people tend to organize into different groups.

However, the community detection problem does not give a precise definition of what a community is. A common definition of a community is that it is densely connected subgraph [17], [18]. In community detection the number of communities is not predefined. The number of possible communities in the network could be large number and we can not analyse all combinations, so we need algorithms to help us to identify potential communities in the network.

Modularity. Comparing the link density of the community by the link density obtained for the same group of nodes randomly connected we could conclude if the community corresponds to the dense subgraph or the structure is created completely random. The modularity is function that measures the randomness of the each partition. With modularity we can compare the communities and decide which one is better.

For the network with N nodes and L links that partitions into n_c communities. Each community has N_c nodes and L_c links. If number of links is larger than the expected number of links between N_c nodes given in the expected node sequence than these nodes may form the community. We calculate the difference between real network connectivity A_{ij} and the expected number of links between nodes if the network is randomly connected, p_{ij}. The p_{ij} can be obtained by randomizing the original network, but keeping the expected degree of each node unchanged, so $p_{ij} = \frac{k_i k_j}{2L}$.

$$M_c = \frac{1}{2L} \sum (A_{ij} - p_{ij}) \tag{1.10}$$

If modularity is positive, the selected nodes may be community as their connectivity is far from random. If M_c is zero, then the connectivity between nodes is random, and if M_c is negative the nodes do not form the community.

The same idea can be generalized to the whole network: The modularity of network partitioned into n_c communities is then defined as:

$$M = \sum_{c=1}^{n} [\frac{L_c}{L} - (\frac{k_c}{2L})^2] \tag{1.11}$$

The higher modularity indicates that nodes are partitioned in better communities. When we put all nodes into only one community $M = 0$, otherwise if each node is community itself $L_c = 0$ and the sum is negative.

Maximum network modularity indicates the best partions. As there are too many partitions, it is not possible to construct all possible partitions and calculate their modularity. For that we need algorithm that could identify the the best partition.

The first algorithm that was proposed for modularity optimization was greedy algorithm. First it assign each node to community, and start with N communities. Then, we try to merge each pair of communities and calculate the moddularity difference ΔM. Then indentify the community pair for which the difference is largest and merge those two communities. This is repeated untill all nodes merge into single community. The best partition is one with largest M.

Louvain algorithm is optimization algorithm with better scalability than gready algorithm, so it can operate on very large networks. Initially each node is assigned to different community, and similar as before we calculate the difference in the modularity if we move the community of one of its neighbours. Then we move the node i to the community such that modularity becomes larger. This is applyed to all nodes untill no further improvement could be made. In the second step we create weighter network whose nodes are communities identified during first step. The weight of the link between communities is the sum of the weights between the links in the communities, and the number of links inside the community is given as weighted self-loop. Then the first and secound steps are repeated, until there is no more change in the modularity, otherwise until we find the maximum, optimal modularity.

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 traditionaly assortative or dissasortative 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. $\rho = \frac{1}{2}\sum_{ij}(A_{ij} - p)g_i g_j$.

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 $p_{11} > p_{12} > p_{22}$.

Stochastic Block Model is model where each node, in given network G, belongs to one of B blocks. Vector $\theta_i = r$ indicates that node i is in block r, while SBM matrix $\{p\}_{BxB}$, specify the probability p_{rs} that nodes from group r are connected to nodes in group s. The SBM model is looking for the most probable model that can reproduce a given network G. Probability of having model parameters θ, p given network G is proportional to likelihood of generating network G, prior of SBM matrix and prior on block assignments:

$$P(\theta, p|G) = P(G|\theta, p)P(p)P(\theta) \tag{1.12}$$

$$P(G|\theta, p) = \prod_{i<j} p_{r_i s_j}^{A_{ij}} (1 - p_{r_i s_j})^{1-A_{ij}} \tag{1.13}$$

where A_{ij} is number of edges between nodes i and j.

Prior on p is modified for core-periphery model such that $P(p) = 3! I_{0<p_{22}<p_{12}<p_{11}<1}$, while prior on θ consists of three parts: probability of having l blocks; given the number of layers probability $P(n|l)$ of having groups of sizes $n_1..n_l$ and the probability $P(\theta|n)$ of having particular assignments of nodes to blocks.

For fitting model in the work [19] authors use Metropolis-within-Gibbs algorithm. The likelihood of SBM model increase with number of blocks and model itself does not define optimal number of communities. Inferring minimum description length (MDL) of the model is one approach to decide which model is more likely.

1.4 Network models

1.4.1 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 $r \leq p$

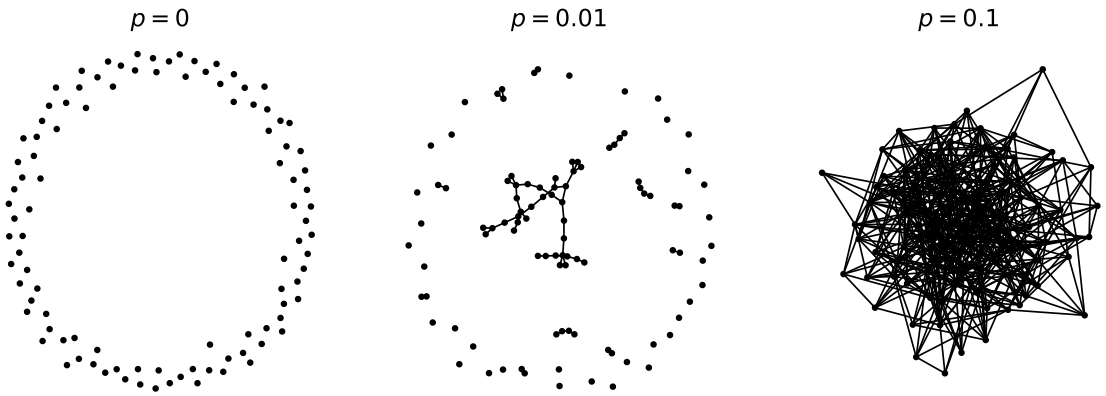

Figure 1.6: ER graph with $N = 100$ nodes and different linking probabilities p.

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

$$\langle L \rangle = \frac{N(N-1)}{2} p \tag{1.14}$$

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:

$$\langle k \rangle = (N-1)p \tag{1.15}$$

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

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{1.16}$$

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 $\langle k \rangle << N$. In this limit, binomial distribution becomes the Poisson, which now depends only on one parameter $\langle k \rangle$

$$p(k) = \frac{1}{k!} e^{-\langle k \rangle} \langle k \rangle^k \tag{1.17}$$

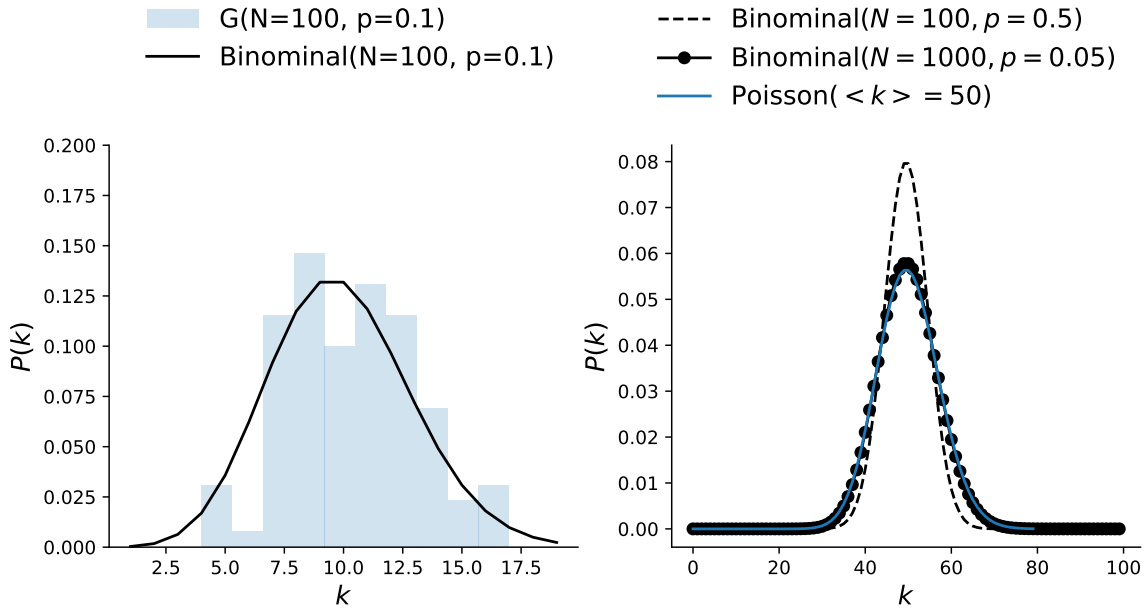

Figure 1.7: Degree distribution of ER graph. Degree distribution of small networks follow binominal. Larger networks are better approximated with Poison distribution, and degree distribution for fixed average degree $< k >$ becomes independent of the network size.

The random graph has a very small **average path length**, it is given as $\langle l \rangle = \frac{lnN}{ln(pN)}$ 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, contrary to real-world networks.

The figure 1.6 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.

1.4.2 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 rewire 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.

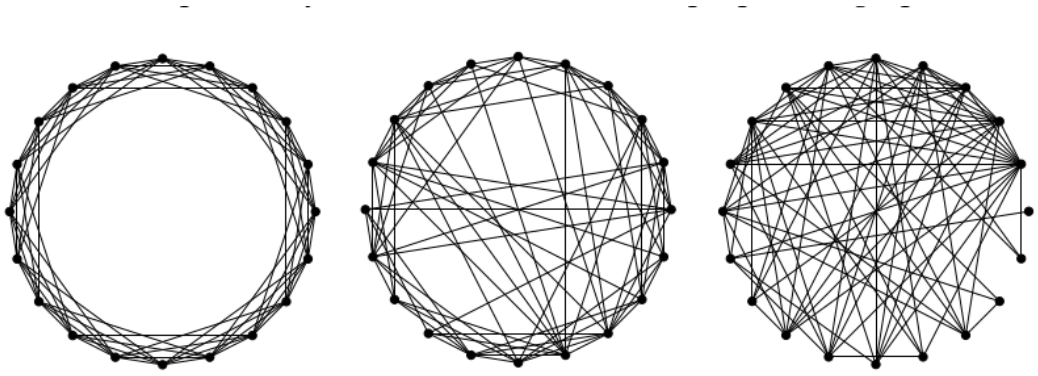

Figure 1.8: Watts and Strogatz graph model creation

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.

1.4.3 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; i 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 Π is then proportional to node degree k:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{1.18}$$

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.

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j} = m\frac{k_i}{m_0 + 2mt} \tag{1.19}$$

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.

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} \tag{1.20}$$

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:

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{1.21}$$

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, $d \sim \frac{lnN}{lnlnN}$. The diameter grows slower than lnN, 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 $C \sim \frac{lnN^2}{N}$. 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 Π. 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.

1.4.4 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:

$$\Pi(k_i) = {k_i}^{\beta} \tag{1.22}$$

The probability that newly added node attaches to node i depends on the existing i-th node degree k_i, and the parameter β. 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 week for having the pure power-law. Instead degree distribution converge to stretched exponential.

1.4.5 Ageing model

To understand how aging can impact the network structure we look into probability dependent on two parameters, nodes degree k and age of node i at the time point t $\tau_i = (t - t_i)$, where t_i is the time when node i is added to the network.

$$\Pi_i(t) \sim k_i \tau_i^{\alpha} \tag{1.23}$$

The parameter α controls the linking probability dependence on the nodes' age if $\alpha = 0$, the ageing of nodes is disregarded.

If $\alpha > 0$ is positive, the older nodes are more likely to create connections. In this regime, the preferential attachment stays present, and the high-degree and older nodes are preferred. For very

high α, each node is connected to the oldest node in the network. The scale-free properties are present; the power-law exponent γ deviates from $\gamma = 3$. It is found that γ ranges between 2 and 3.

When α is negative, ageing overcomes the role of preferential attachment, and scale-free properties are lost. For significant negative α network becomes a chain; the youngest nodes are those who get connected.

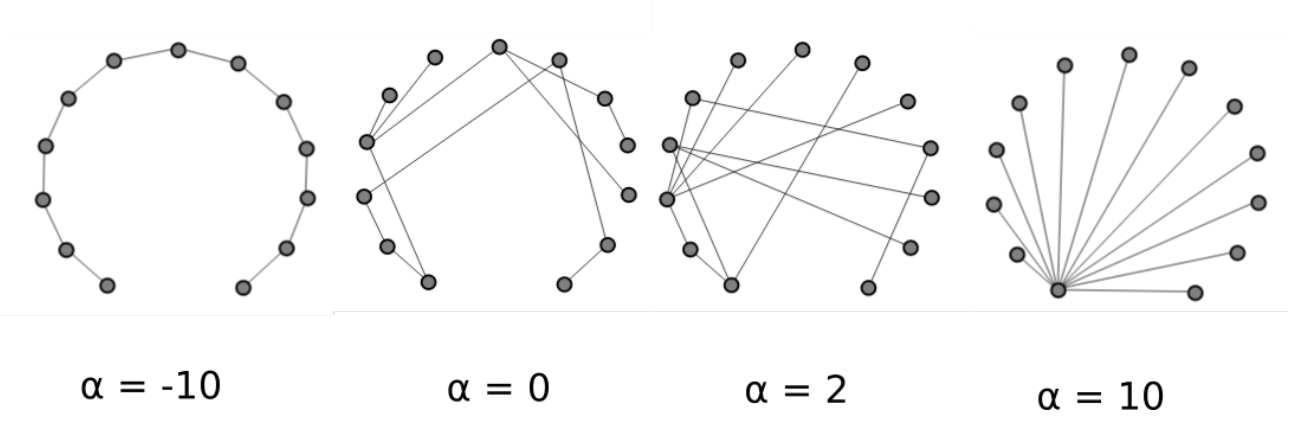

Figure 1.9: Aging model

In the general ageing model, the non-linearity on the node degree is introduced, so this model has two tunable parameters α and β. The probability that a link is created between the new node and the existing node is defined as

$$\Pi_i(t) \sim k_i(t)^{\beta} \tau_i^{\alpha} \tag{1.24}$$

As before, depending on model parameters network evolves to different structures [20].

- For example if we fix $\beta = 1$ and $\alpha = 0$ generated networks are scale-free; degree distribution is $P(k) \sim k^{-\gamma}$ with $\gamma = 3$.
- 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.
- For $\alpha < \alpha^*$ and near critical line $\beta(\alpha^*)$ degree distribution has **stretched exponential** shape

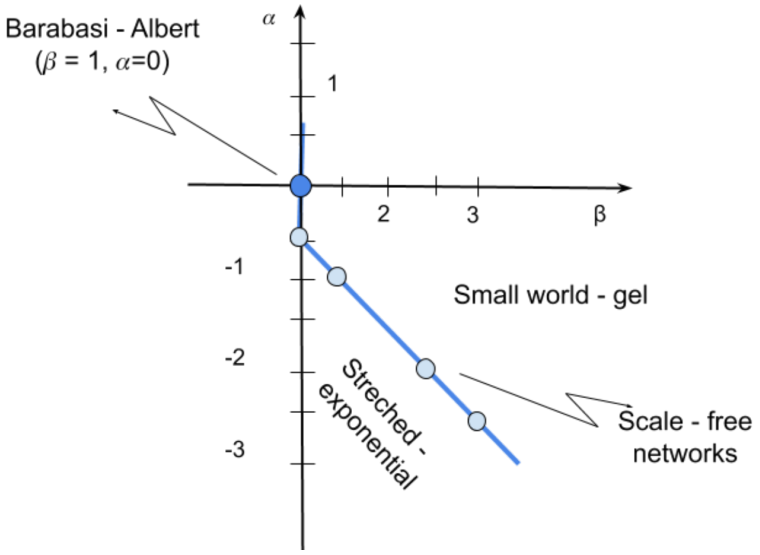

Figure 1.10: Phase diagram of aging network model

1.4.6 Stochastic block model

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: $g_i \in \{1, 2..k\}$, 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 1.11 [17] shows how the model matrix corresponds to resulting networks with two communities. First, for the assortative network (1.11 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, (1.11 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 (1.11 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 (1.11 d) because all probabilities inside and between groups are equal.

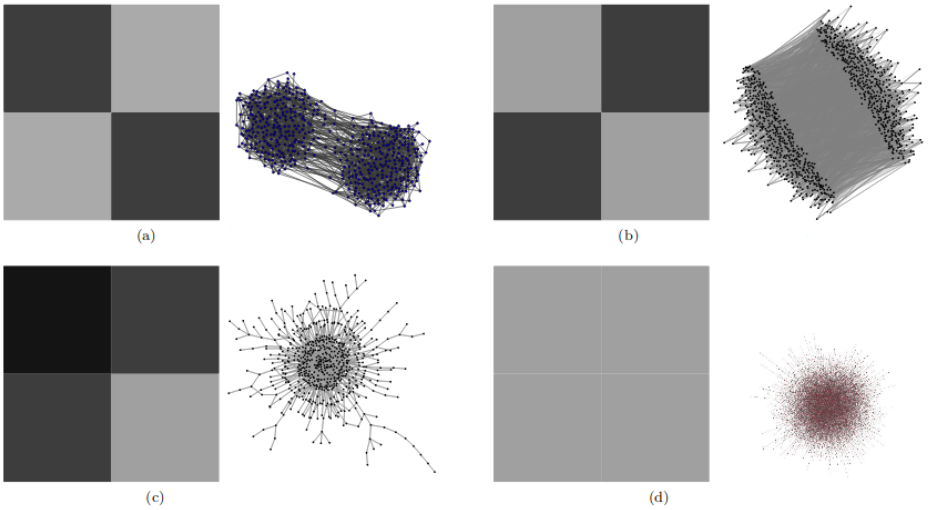

Figure 1.11: Stochastic Block model for different networks structures. (a) assortative. (b) dissortative. (c) core-periphery. (d) Erdos Renyi random graph.

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 [21].

1.5 The probability distributions

The shape of degree distribution is important for getting the first insight into the characteristics of the complex network. When nodes are generated at random and any two nodes are linked with the same probability p, we expect the binomial distribution, or for larger networks it is Poisson distribution $P(k) = \frac{1}{k!}e^{-\langle k\rangle}\langle k\rangle^k$, where $\langle k\rangle = Np$. A different approach is to add one node and connect it randomly to the network at each time step. The obtained network then has the exponential degree distribution $P(k) = e^{-\lambda k}$. These are exponentially bounded distributions, meaning they decay exponentially or faster for the large values.

On the other hand, heavy-tailed distributions decay slower than exponential, and the events for large values are rare but still possible. For example, in the preferential attachment model, degree distribution emerges to the power law. Also, many empirical data exhibit the heavy-tailed distribution. Even if they look like a power law, after statistical analysis, it may be concluded that the data deviate from the power law and could be equally good or even better fitted with some other distribution. Commonly used alternative distributions are log-normal distribution, stretched-exponential or power-law with an exponential cutoff.

This section gives an overview of relevant distributions and methods for fitting data and testing

the quality of the performed fit.

1.5.1 The properties of distributions

Power-law The power-law distribution is defined as

$$p(k) = Ck^{-\gamma} \tag{1.25}$$

where parameter γ is an exponent of the power-law distribution while the C is the normalising constant.

The distribution can take discrete and continuous values, which is defined for positive values $k > 0$, so there is a lower bound to the power-law function k_{min}. For the discrete case $C = 1/\zeta(\gamma, k_{min})$, while in the continuous case $C = (\gamma - 1)k_{min}^{\gamma-1}$.

The power-law distribution is called scale-free distribution. If we scale the value k for the factor 2 the ratio of $p(x)/p(2x)$ is constant and does not depend on the k. We'll find that these criteria are not satisfied by any other distribution.

$$\frac{p(k)}{p(2k)} = \frac{Ak^{-\gamma}}{A(2k)^{-\gamma}} = 2^{\gamma} \tag{1.26}$$

The scale-free function is defined as $p(bx) = g(b)p(x)$. The solution of this equation is $p(x) = p(1)x^{-\gamma}$, where $\gamma = -p(1)/p^{'}(1)$ lead us to conclusion that if function is self-similar it has to be power-law.

Lognormal distribution. The variable x has the lognormal distribution if the random variable $y = ln(x)$ is distributed as normal distribution.

$$f(y) = \frac{1}{2\pi\sigma} e^{-(y-\mu)^2/2\sigma^2} \tag{1.27}$$

where μ is the mean, and σ is the standard deviation. The density distribution of the log-normal distribution is defined as

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-(log(x)-\mu)^2/2\sigma^2} \tag{1.28}$$

The lognormal distribution has finite mean $e^{\mu+1/2\sigma^2}$, and the variance $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$. [22]. Despite the finite moments, the log-normal distribution can be similar to the power-law distribution. If the variance is large, then the probability function on the log-log plot appears linear for a large range of values.

Using the **multiplicative processes**, we can generate the log-normal distribution. The log-normal distribution is generated by processes that economist Gibrat called the law of proportionate effect. If we start from the organism of size S_0. At each time step, the organism may grow or shrink according to the random variable ϵ,

$$S_t = \epsilon_t S_{t-1} \tag{1.29}$$

When the system's state at time t is proportional to the state at the previous time step, we have the multiplicative process. The ϵ is a proportionality constant that can change over time. The current state depends only on the initial size S_0 and the ϵ variables.:

$$S_t = \epsilon_t S_{t-1} = \epsilon_t \epsilon_{t-1} ... \epsilon_2 \epsilon_1 S_0 \tag{1.30}$$

If ϵ_t is drawn from the log-normal distribution, then S_t also follows log-normal, as the product of log-normal distributions is again log-normal. Still, the ϵ distribution does not determine the distribution of the S_t. Taking the logarithm of the equation:

$$ln(S_t) = ln(S_0) + \sum_{i=0}^{t} ln(\epsilon_i) \tag{1.31}$$

The sum of the logarithms of the ϵ_t, according to the Central Limit Theorem (CLT), follows the normal distribution. The CLT states that the sum of identically distributed random variables with finite variance converges to the normal distribution. If $ln(S_t)$ is normally distributed, then S_t follows the log-normal distribution.

The multiplicative processes generate the log-normal distribution. Introducing threshold in the multiplicative process leads to the power law. For example, in the Champernowne model, individuals are divided into classes according to their income. The minimum income is m. People between incomes m and γm are in the first class, and the second class are people with incomes between γm and $\gamma^2 m$. The individuals can change their class, so it is described as a multiplicative process, but with a threshold, as income can not be lower than m. If we fix $\gamma = 2$, and consider that with probability $p_{i,i-1} = 2/3$ the change is from higher to lower class. In contrast, with probability, $p_{i,i+1} = 1/3$ individual goes to higher class. In this process, the distribution of incomes emerges to the power-law distribution.

Power law with exponential cutoff. The density function has the following form

$$p(k) = Ck^{-\gamma}e^{-\lambda k} \tag{1.32}$$

where $k > 0$ and $\gamma > 0$. This function combines the power-law and exponential terms responsible for an exponentially bounded tail. Taking the logarithm $ln(p(k)) = lnC - \gamma lnk - \lambda k$, when $k << 1/\lambda$ the second term dominates, so distribution follows te power-law, with exponent γ. Otherwise, the λx term dominates, resulting in an exponential cutoff for high values.

Streched exponential The stretched exponential distribution is defined as:

$$p(k) = ck^{\beta-1}e^{-(\lambda k)^{\beta}} \tag{1.33}$$

the parameter β is stretching exponent determining the properties of the function $p(k)$. For $\beta = 1$, the function is exponential. For $\beta < 1$ it is hard to distinguish the distribution from the power law. We have a compressed exponential function for $\beta > 1$, so k varies in the narrow range.

1.5.2 Plotting the distributions

The first step in analysing the empirical data is to create the frequency plot or histogram. Data are binned in equal intervals, and the number of data points within the interval are plotted. It is hard to determine whether the distribution is exponential or power law when plotting heavy-tailed distributions. If data are from power law distribution on the double logarithmic scale, they will look linear:

$$logp(k) = \gamma log(k) + c \tag{1.34}$$

On the log-log scale, we can notice that in the tail of the distribution data are noise. As the size of the bins is constant, the bins' density for large values also becomes large. To avoid the fluctuations in the tail, we can use logarithmic binning. The noise is reduced by dividing the x axis into n bins $b_n = c^n$, so the following bin is wider than the previous one. For the base c we can choose any value $c > 1$. Similarly, the binning can take the following form $b_n = k_0 \exp(cn)$, where k_0 is the

minimum data point, while the c is the arbitrary base. All data points between values $[b_n, b_{n+1})$ are represented with one point $p(k_n) = N_n/b_n$, where N_n is number of nodes found in the bin b_n and $k_n = \sum_i k_i/N_n$ is average degree of the nodes in the bin b_n. By averaging over bins in the tail, noise in the tail of the distribution is reduced. Still, no matter how bin size is chosen, the information about original data points is lost, especially in the distribution tail where bins are larger and include more samples. Figure 1.12 shows how different distributions look like on linear (first column) and log-log scale (second column).

Instead of plotting the probability distribution it is possible to calculate the cumulative distribution, defined as $P(k) = \int_k^\infty p(k^{'})dk^{'}$ for continuous function or as $P(k) = \sum_{k^{'}=1}^{k} p(k^{'})$ for the discrete function. For example the CDF function for power law is also power-law function but with exponent $\gamma - 1$: $P(k) = k^{-(\gamma-1)}$. Note that for cumulative distribution, it is not necessary to use log-binning.

1.5.3 Estimating the distribution parameters

The maximum likelihood estimation(MLE) is a method where we consider that data comes from a particular distribution, so we want to maximise the likelihood of the data to find the distribution parameters. For given set of i.i.d. observations $x_1, x_2, ...x_n$, sampled from the distribution $p(x)$ we can define the likelihood function [23]. The likelihood function tells us how likely it is to have the given data if the distribution parameters are θ.

$$L(\theta|x_1, ...x_n) = \prod_{i=1}^{i=n} p(x_i|\theta) \tag{1.35}$$

The parameter that maximize the likelihood function is $\theta_{max} \in argmax L(\theta|x_1, ...x_n)$.

We can solve the equation and derive the expression for maximum likelihood parameters. The parameters can be obtained with numerical optimisation for distributions where an analytical solution is unavailable. In practice is much easier to work with logarithm of the likelihood, $log(L) = \sum_{i=1}^{i=N} p(\theta|x_i)$, as the product changes to summation. For the power-law distribution, the exponent is calculated as $\gamma = 1 + n[\sum ln\frac{k_i}{k_{min}}]^{-1}$. For discrete distribution solution may be obtained optimizing the log-likelihood function $log(L) = log \prod_{i=1}^{n} \frac{k_i^{-\gamma}}{\zeta(\gamma, k_{min})}$.

We can use the MLE method to fit any distribution to the data. Even if obtained distribution looks like a power law, and some parameters are estimated, it does not have to be that data are truly from the power-law distribution. With the MLE method alone, it is impossible to distinguish between different distributions, and we do not know how accurate the obtained results are. To determine the quality of the fit, we need to use another statistical method called the **goodness-of-the-fit** test. The main idea is based on calculating the distance between distributions of empirical data and the model using Kolmogorov-Smirnov statistics. The Kolmogorov Smirnov statistics is the maximum distance between the CDF of the data and the fitted model.

$$D = max|S(x) - P(x)| \tag{1.36}$$

First, we fit empirical data to get model parameters and calculate the KS statistics of this fit. Then, many synthetic data sets are generated with model optimised model parameters. Then each synthetic data set is fitted, and KS statistics are obtained relative to its model. From there, we can calculate **p-value**, the fraction of times that KS-statistics in synthetic distributions is larger than in empirical data.

If $p - value < 0.1$, we reject the hypothesis that this distribution describes the empirical data. Otherwise, the model can not be rejected. Failing to reject the hypothesis does not mean the model is

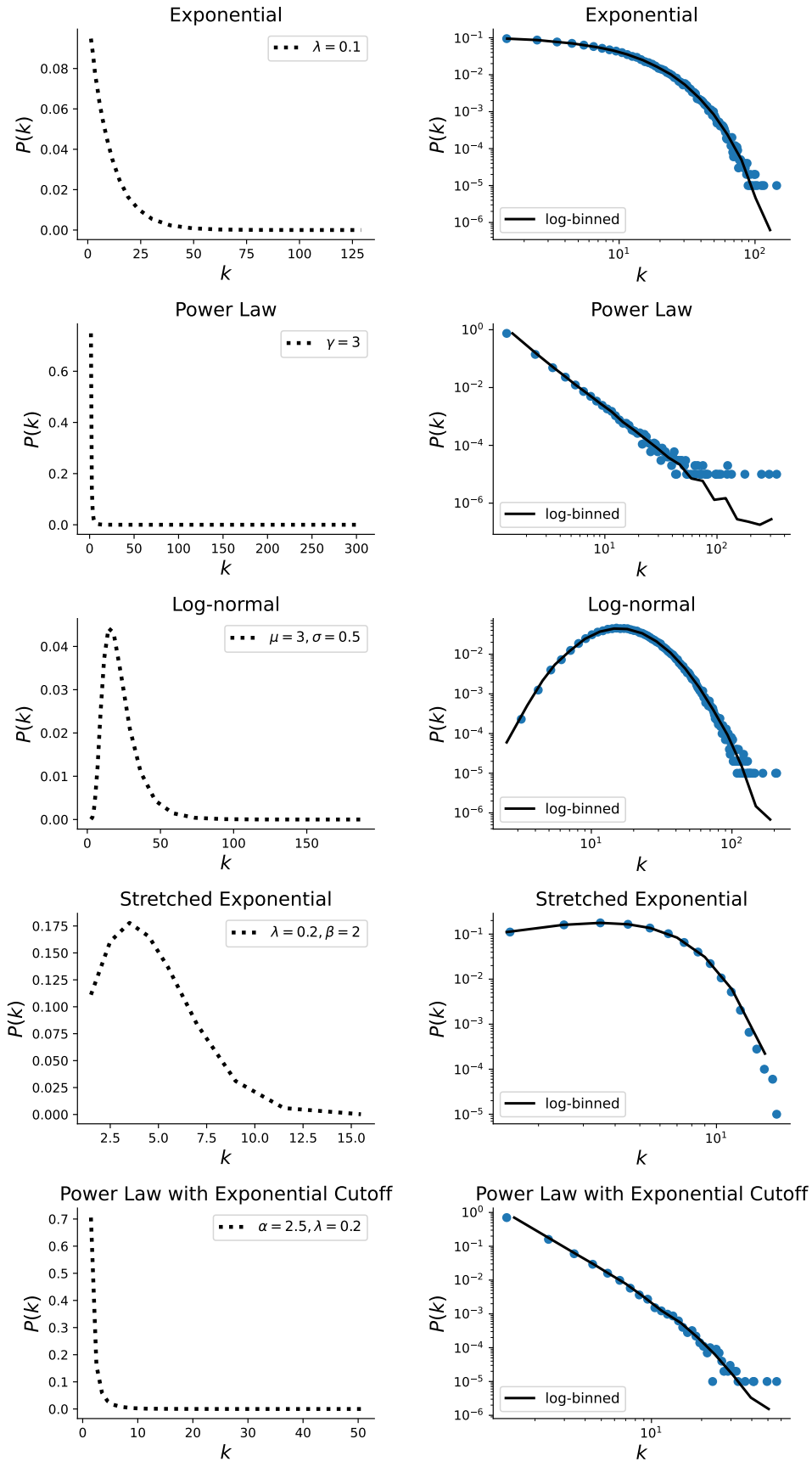

Figure 1.12: Probability distributions on a linear and double logarithmic scale.

a correct distribution for the data. Other distributions might fit the data equally good or even better.

To have an accurate p-value, we need a large sample. For a small number of synthetic distributions, it is possible to have a high p-value, even if the distribution is the wrong model for the data. Finally, we need to be confident in obtained results. The same procedure can be repeated for different distributions. If the p-value for the power law is high, while for alternative distribution, it is low, we can conclude that the power law is a more probable fit.

Another method, the **likelihood ratio test**, allows us to compare two distributions directly. The distribution with a higher likelihood under empirical data is a better fit. We can calculate the likelihood ratio, or it is easier to obtain the likelihood ratio's logarithm because its sign determines which distribution is a better fit. For given two distributions $p_1(x)$and $p_2(x)$.

The likelihoods are defined as $L_1 = \prod_{i=1}^{n} p_1(x)$ and $L_2 = \prod_{i=1}^{n} p_2(x)$, or the ratio of likelihoods as $R = \frac{L_1}{L_2} = \prod_{i=1}^{n} \frac{p_1(x)}{p_2(x)}$. Taking the logarithm, we obtain the log-likelihood ratio

$$\mathcal{R} = \sum_{i=1}^{n} [logp_1(x_i) - logp_2(x_i)] \tag{1.37}$$

As data x_i are independent, by central limit theorem, their sum $\mathcal{R}$ becomes normally distributed, with expected variance σ^2. We can approximate the variance as

$$\sigma^2 = \frac{1}{n} \sum_{1}^{n} [(l_i - l_i) - (< l >^{(1)} - < l >^{(2)})]$$

When $R > 0$ the first distribution is a better fit to data, and when $R < 0$, the other one should be chosen. When $R = 0$, it is not possible to distinguish between two distributions. The sign of R is not enough criteria to conclude which distribution is a better fit, and it is a random variable subject to statistical fluctuations. We need a log-likelihood ratio that is sufficiently positive or negative to ensure that its sign does not result from fluctuations.

If we are suspected that the expectation value of the log-likelihood ratio is zero, the observed sign of is simply the product of fluctuations and can not be trusted. The probability that the measured log-likelihood ratio has a magnitude as large or larger than the observed value R is given as

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}} \int_{-\infty}^{-|\mathcal{R}|} e^{-x^2/2n\sigma^2} dx + \int_{|\mathcal{R}|}^{\infty} e^{-x^2/2n\sigma^2} dx \tag{1.38}$$

Here we use the standard two-tail hypothesis test, assuming that the null hypothesis is $R = 0$. If the p-value is larger than a threshold, the R sign is unreliable, and the test does not favour any distribution. If p is small, $p < 0.1$ then it is unlikely that the observed sign is obtained by chance, so we reject the null hypothesis that $R = 0$.

1.6 Fractal analysis

The approach to studying complex systems is detecting the time series of selected variables. In complex systems, the periodic behaviour of time series is not limited to one or two characteristic frequencies. They extend over a wide spectrum and fluctuations on many time scales and broad distributions. In these cases, the system's dynamics is characterised by scaling laws, which are valid over a wide range of time scales or frequencies. When only one scaling exponent describes the system dynamics, the time series is monofractal. On the other hand, we deal with multifractal time series.

Rescaling time t by a factor a may require rescaling the time-series values $x(t)$ by a factor a^H; then, we have the self-similarity. The Hurst exponent, H characterise the type of self-affinity.

$$x(t) = a^H x(at)$$

1.6.1 Long and Short-term correlations

The time series are persistent; large values usually follow a large value. Considering the increments $\delta x_i = x_i - x_{i-1}$, of self-affine series $i = 1, .., N$, with N values measured equidistant in time, so δx_i can be either persistent, independent or anti-persistent. For the random walk with $H = 0.5$ the increments are independent. For stationary data with constant mean and standard deviation, the auto-covariance function can determine the degree of persistence.

$$C(s) = \langle \Delta x_i \Delta x_{i+s} \rangle = \frac{1}{N-s} \sum_{i=1}^{N-s} \Delta x_i \Delta x_{i+s} \tag{1.39}$$

If the data are uncorrelated, the $C(s) = 0$. Short-range correlations are described by $C(s)$ declining exponentially

$$C(s) = exp(-s/t_c)$$

such behaviour is typical for increments generated by an auto-regressive process

$$\Delta x_i = c \Delta x_{i-1} + \epsilon_i$$

with random uncorrelated offsets ϵ_i and $c = exp(-1/t_c)$.

For long-range correlations, $\int C(s)$ diverges in the limit for long series. In practice, this means that we can not define the characteristic time because it increases with N. Contrary to short-range correlations, the correlation function decline as power-law

$$C(s) = s^{-\gamma}$$

Fourier filtering techniques can model this type of behaviour. Long-term correlated behaviour of Δx_i leads to self-affine scaling behaviour characterised by Hurst exponent $H = 1 - \gamma/2$.

A direct calculation of the $C(s)$ is difficult due to present noise in the data and non-stationarity. Non-stationarities make the definition of $C(s)$ problematic because its average is not well defined. Also, $C(s)$ fluctuates around zero on large scales s, so it is impossible to obtain the correct correlation exponent γ. Instead of calculating $C(s)$, we can calculate the Hurst exponent H.

1.6.2 Rescaled range analysis

Hurst proposed method called the **rescaled range analysis** R/S. It begins with splitting the time series x_i into non overlapping segments ν of the size s, having $N_s = int(N/s)$ segments. Then is calculated the profile in each segment is.

$$Y_\nu(j) = \sum_{i=1}^{j} (x_{\nu s+i} - \langle x_{\nu s+i} \rangle_s)$$

Substracting the averages, constant trends in the data are eliminated. The differences between minimum and maximum value and the standard deviation in each segment are calculated as: $R_\nu(s) = maxY_\nu(j) - minY_\nu(j)$, $S_\nu(s) = \sqrt{\frac{1}{s} \sum Y_\nu^2(j)}$

Finally, the rescaled range is averaged over all segments to obtain the fluctuation function F(s).

$$F_{RS}(s) = \frac{1}{N_s} \sum \frac{R_\nu(s)}{S_\nu(s)} \sim s^H$$

, where the H is the Hurst exponent. Values $H < 1/2$ indicate long-term anti-correlated data while $H > 1/2$ long-term positively correlated data.

1.6.3 Fluctuation analysis

The **fluctuation analysis** is based on the random walk theory. For given time series $\{x_i\}$ with length N, we first define the global profile in the form of cumulative sum, equation 1.40, where $\langle x \rangle$ represents the average of the time series.

$$Y(j) = \sum_{i=0}^{j} (x_i - \langle x \rangle), \quad j = 1, ..., N \tag{1.40}$$

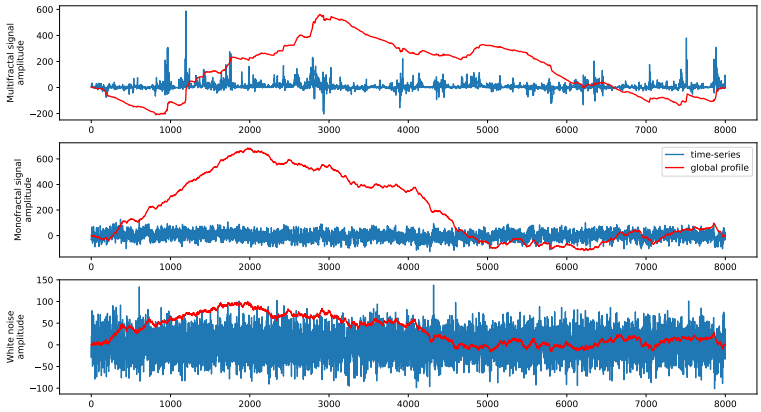

Figure 1.13: Multifractal, monofractal and whitenoise signals.

The profile of the signal Y is divided into $N_s = int(N/s)$ non overlapping segments of length s. If N is not divisible with s the last segment will be shorter. That is handled by doing the same division from the opposite side of the time series, giving us $2N_s$ segments. Then we calculate the fluctuations in each segment $F^2(\nu, s)$ and finally, average overall subsequences, obtaining the mean fluctuation. From the scaling of the function, we can determine the Hurst exponent.

$$F_2(s) = [\frac{1}{2N_s} \sum F^2(\nu, s)]^{1/2} \sim s^H \tag{1.41}$$

Several methods are proposed for calculating the fluctuating function $F^2(\nu, s)$:

- The most straightforward way to calculate the fluctuations is to consider the difference in the values at the endpoints of each segment. It is same as eliminating the linear trend from each segment.

$$F^2(\nu, s) = [Y(\nu s) - Y((\nu + 1)s)]^2$$

- The trends present in the time series does not have to be linear. When dealing with non-stationary time series, removing the polynomial trend within each segment is necessary by least-square fitting. The method is called detrended fluctuation analysis (DFA). From each segment ν, local trend $p^m_{\nu,s}$ - polynomial of order m - should be eliminated, and the variance $F^2(\nu, s)$ of a detrended signal is calculated as in equation

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \left[Y(j) - p^m_{\nu,s}(j) \right]^2 \tag{1.42}$$

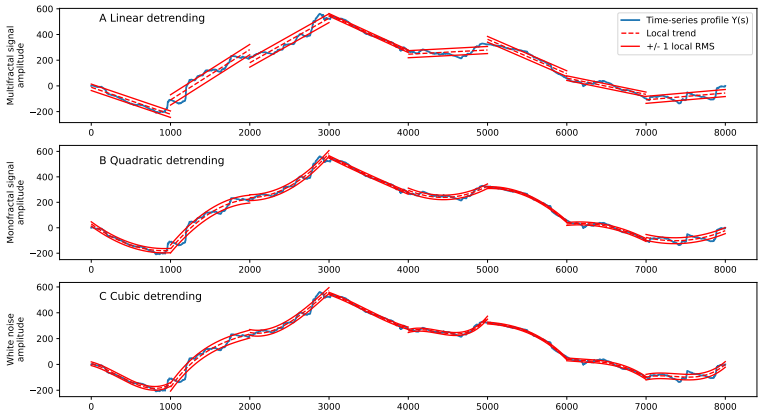

Figure 1.14: Detrending the signal, for the segments of length $s = 1000$.

1.6.4 Multifractality of the signals

The scaling behaviour in many data may be more complicated, and different scaling exponents can be found for many interwoven subsets of the time series, representing multifractal. The multifractality may come from the broad probability distribution of the time series values. In this case, the multifractal properties can not be destroyed with shuffling time series. The source of multifractality may be from different small and large fluctuations correlations. In this case, the probability density function of the values can be regular distribution with finite moments, and the corresponding shuffled series will exhibit non-multifractal scaling as correlations are destroyed with the shuffling procedure. When both kinds of multifractality are present, the shuffled time series will show weaker multifractality.

The multifractal analysis will reveal higher-order correlations. Multifractal scaling can be observed if the scaling behaviour of small and large fluctuations is different. Multifractal detrended fluctuation analysis (MFDFA) is used [24, 25] to estimate multifractal Hurst exponent H(q).

$$F_q(s) = \left\{ \frac{1}{2N_s} \sum_{\nu}^{2N_s} \left[F^2(\nu, s) \right]^{\frac{q}{2}} \right\}^{\frac{1}{q}}, q \neq 0$$

The MFDFA for $q = 2$ is equivalent DFA method. The value of $H(0)$, which corresponds to the limit $H(q), q - 0$, cannot be determined directly because of the exponent diverge. Instead, the

logarithmic averaging procedure has to be considered.

$$F_0(s) = \exp\left\{\frac{1}{4N_s}\sum_{\nu}^{2N_s} ln\left[F^2(\nu, s)\right]\right\}, q = 0 \tag{1.43}$$

The fluctuating function scales as power-law $F_q(s) \sim s^{H(q)}$ and the analysis of log-log plots $F_q(s)$ gives us an estimate of multifractal Hurst exponent $H(q)$.

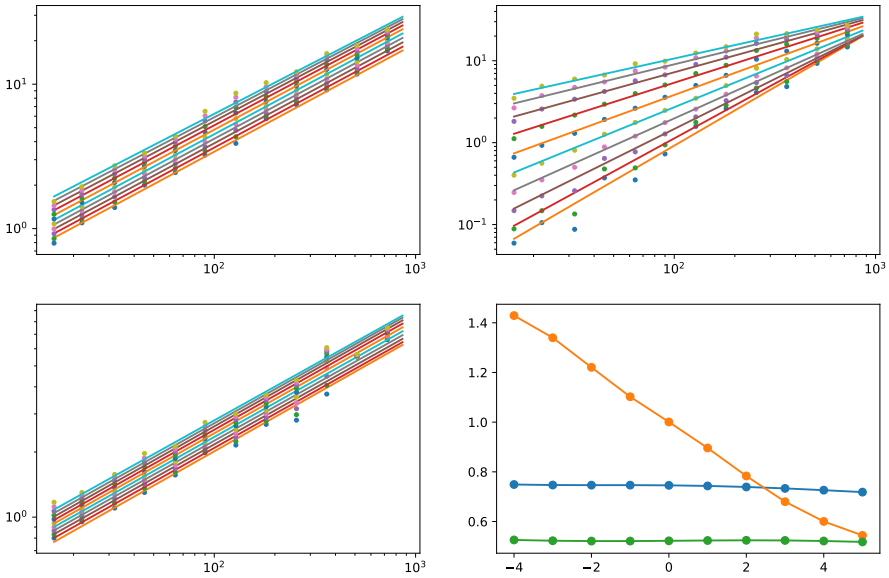

Figure 1.15: Multifractal, monofractal and whitenoise signals.

For the monofractal time series, H(q) is independent of q, meaning that scaling is identical for all segments, and averaging fluctuations gives identical scaling for all values of q. If small and large changes scale differently, h(q) will be dependent on q. Positive values of q, segments with large variance are dominant in the Fq(s), so positive q describes segments with large fluctuations. The negative values of q, H(q) describe the scaling of the segments with small fluctuations.

1.7 Dynamical reputation model

Consider a system where each component has an activity pattern that could be mapped to the discrete signal, representing the moments when the event happened, such as the activity pattern when users are sending an email or communicating, sharing opinions and information within the community. Users' behaviour directly influences their position in the community, which is measured through reputation. The trust among users depends on the amount of interaction between them, which means the trust changes over time. The computational model needs to capture the dynamical property of the trust. Furthermore, the important property of trust is that it is easier lost than gained; the frequency of interaction also matters. The trust between users who interact frequently should increase faster than between users who rarely interact.

With Dynamic Interaction Based Reputation Model (DIBRM) [26] we can quantify the user reputation R_n after each interaction using equation 1.44, where n is number of interaction $n \in 1, N$.

$$R_n = R_{n-1}\beta^{\Delta_n} + I_n \tag{1.44}$$

The first part of the equation considers the reputation value after previous interaction R_{n-1}, weighted with coefficient β_n^{Δ}. Depending on the frequency of the interaction, reputation will rise or decay. Parameter β ranges from $0 < \beta < 1$ is forgetting factor. The Δ_n measures time between two interactions t_n and t_{n-1}:

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a} \tag{1.45}$$

where t_a is the characteristic time window of interaction. In the second part of the equation, I_n is the reputation gained within each interaction. The basic value of each interaction is given as I_{bn}, and the parameter α is the weight of the cumulative part.

$$I_n = I_{b_n}(1 + \alpha(1 - \frac{1}{A_n + 1})) \tag{1.46}$$

When $\Delta_n < 1$, a user is frequently active, meaning that the time between two interactions is less than the characteristic time window. The number of sequential activities A_n increases by 1. On the other hand, when $\Delta_n > 1$ is large, the reputation decays, while the number of activities resets to $A_n = 1$.

If a user is frequently active, we can record the reputation after each day. On the other hand, if $t_n - t_{n-1} > 1$day we need to interpolate the reputation values for each day between two interactions, $t_{n-1} < t_d < t_n$. To do that we consider that due to inactivity reputation will only decay, so it could be calculated as $R_d = R_{n-1}\beta^{\Delta_d}$, where $\Delta_d = (t_d - t_{n-1})/t_a$.

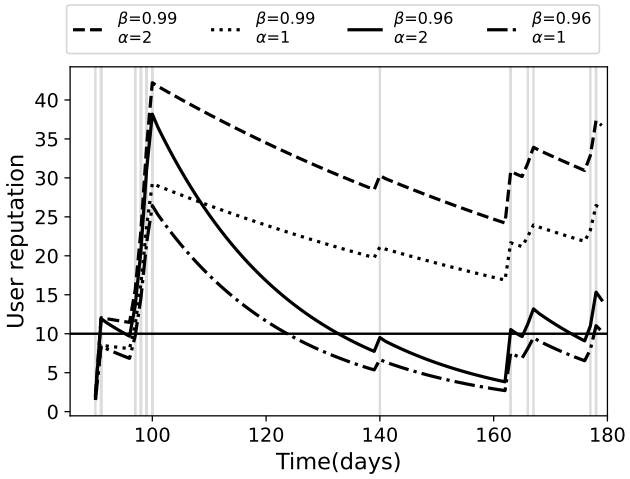

Figure 1.16: User reputation.

For example, if we set the characteristic window size and basic value of interaction to $t_a = 1day$, $I_{bn} = 1$, we can analyze the influence of the parameters α and β on the user reputation. Lower beta and alpha values lead to faster reputation decline as shown in figure 1.16. With lower β the reputation may quickly drop close to the reputation threshold, under which we don't consider the user as active. In contrast, with larger values of β, reputation stays high even if a user is not active for a larger period. The parameter α is the most important influence on burst behaviour, where larger α leads to higher reputation values.

When a user becomes inactive, its reputation starts to decline, and when it drops below the reputation threshold user does not have any influence on the community. We can approximate the dependence of parameter β and time δt needed for reputation to reach this level as $\beta = (\frac{R_0}{R_i})^{\frac{t_a}{\delta t}}$. In the examples on figure, parameter $t_a = 1$day, while we vary different starting reputation levels R_n.

For β values below 0.96, the decay is fast, and within two to four months of inactivity, even high reputation values are reduced below the threshold. On the other hand, with values of β, the decay process is more differentiated, and the high reputation becomes harder to lose, surviving up to a year of inactivity. For β equal to 0.96, it takes a month for the reputation based on 5 interactions to decay and around 5 months for the high reputation based on 500 or 1000 interactions to decay below the threshold.

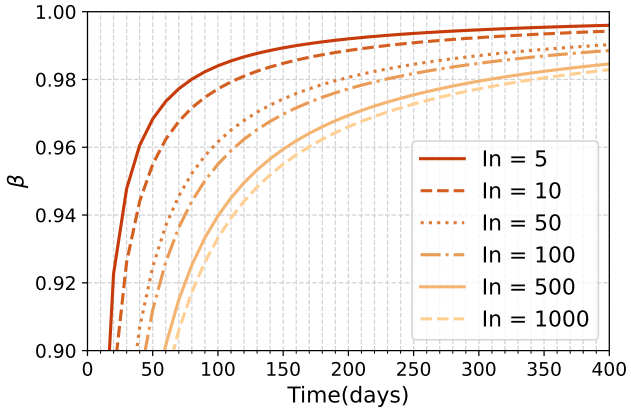

Figure 1.17: Reputation decay

In this model, the user's reputation changes continuously through time, decreases when the user is inactive and grows with frequent and constant user contribution. The highest growth of a user's reputation is found through bursts of activity followed by a short period of inactivity. With model parameters, $I_{bn}, t_a, \alpha, \beta$, the dynamic of user reputation may be controlled and adapted to different communities. If the community has its reputation model, we can also fit the model parameters to mimic the actual reputation dynamic. In this thesis, the DIBRM model is used to analyze the reputation of users in StackExchange communities. Chapter 5 gives details about model parameters choice.

Chapter 2

Driving signals

The complex networks grow by adding new nodes, and growing network models consider growth constant over time. This approximation is sufficient for explaining how properties of complex networks can emerge; for example, we find scaling of degree distribution in the Barabasi-Albert model, such as in real systems. Models mainly focus on linking rules and their influence on the topology of complex networks.

Still, the growth of real systems changes over time. In online social networks, new users join daily, and the users' activity might have bursty nature. We can consider a co-authorship network, where links are created between scientists when they publish a paper. The dynamic of real networks can be complex and highly influenced by non-linear signals. The growth signal, the number of new nodes in each time step, has cycles and trends. Circadian cycles are directly reflected in growth signals, and we also find long-range correlations and multifractal properties.

In this chapter, we explain the properties of growth signals, both real and computer-generated. We analyze networks created with a growing network model where the interplay between ageing and preferential attachment shapes their structure. We are interested in incorporating non-constant growth signals into the model and measuring their impact on the complex networks. Differences between networks with the same number of nodes and links can be observed by analyzing connectivity patterns. Figure 2.1 describes our goals.

2.1 Aging network model with growth signal

To enable nonlinear network growth in the number of nodes, we need to adapt the existing models such that at each time step, we can add $M \geq 1$ new nodes that make $L \geq 1$ links with existing nodes in the network. The master equation N_k, k degree nodes can be written as:

$$\partial_t N_k = \sum_{j=1}^{M(t)} r_{k-j \longrightarrow k} N_{k-j} - \sum_{j=1}^{M(t)} r_{k \longrightarrow k+j} N_k + M(t)\delta_{k,L}. \tag{2.1}$$

At each time step we add $M(t)$ nodes with L links. As multiply links between two nodes are not allowed, we'll get $M(t)$ new nodes with degree L, which describes the third term in the equation. Old nodes can increase their degree from 1 to $M(t)$, as different new nodes can choose the same node.

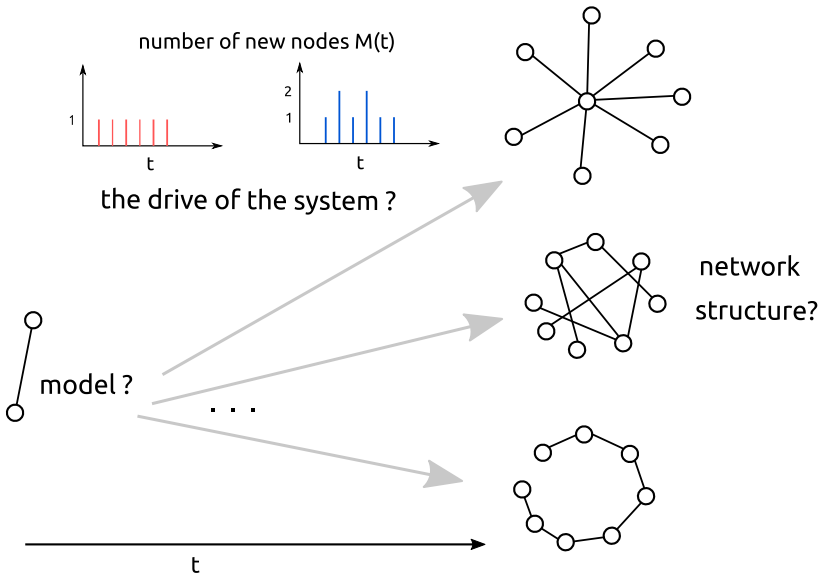

Figure 2.1: The open question is how nonlinear signals in combination with network model influence the structure of the network. Under what circumstances networks have the scale-free, hub-spoke or chain structure.

The first term in the equation describes nodes with degree $k \in \{k - M(t), \ldots, k - 1\}$ that getting degree k, while in second term nodes with degree k entering degree $k \in \{k + 1, \ldots, k + M(t)\}$. The quantities $r_{k-j \longrightarrow k}$ and $r_{k \longrightarrow k+j}$ are the rates that express the transitions of a node from class with degree $k - j$ to one with degree k and from class with degree k to class with degree $k + j$ respectively.

For model we choose, aging model where linking probability depends on network degree k and its age τ, $\Pi_i(t) \sim k_i(t)^\beta \tau_i^\alpha$. With this linking probability the master equation was solved for $M(t) = const. = 1$, using approach [27]. When $M(t)$ is the correlated function, the equation is not solvable analytically. Instead, we use numerical simulations to study the influence of the signal $M(t)$ on the network structure. When we add only one link per node $L = 1$, networks are uncorrelated trees. To obtain the clustered structures, we need to use $L > 1$; so each new node can create more than one link. Finally we focus on the aging model parameters $-\infty < \alpha \leq -1$ and $\beta \geq 1$. We expect critical line $\beta(\alpha^*)$ where scale-free networks can be found. Under critical line, networks have stretched exponential degree distribution, and for large β small-world networks are present.

Finally, we need to define what the time series of new nodes are. We focus on the growth of two real systems, the **TECH** [28] community in the Meetup website and on two months of **MySpace** [29] social network.

2.1.1 Time-series from real systems

MySpace signal is the number of new members who appear for the first time in the data. The dynamics of the Meetup website happen on different scales than on Meetup. Here, the time step is one minute. The MySpace signal has $T = 3162$ steps, with $N = 10000$ members. To describe the properties of the signal, we use Multifractal detrended analysis and calculate the Hurst exponent on different scales, showing the right pane of the figure, 2.2. It is multifractal $q < 0$, and becomes constant for $q > 0$, it has long-range correlations as $H(q = 2) = 0.6$. My Space signal has cycles characteristic of the human circadian rhythm, figure 2.2. If we randomize the MySpace signal, we find that we can easily destroy trends and cycles. The randomization is done with the reshuffling procedure, where we

keep number where we keep the number of nodes, length and the mean value of the signal. The inset of the original and randomized signals show the time series' global profile; we find that trends are destroyed. Also, randomized MySpace signal does not have long-range correlations anymore, Hurst exponent indicates short-range correlations $H = 0.5$, and the signal becomes monofractal.

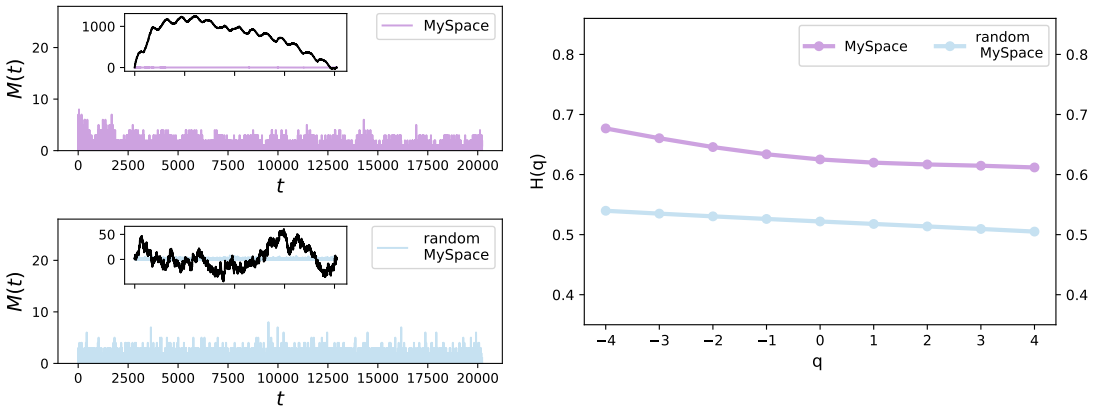

Figure 2.2: MySpace signal, the random MySpace signal (left pane) and the dependence of multi fractal Hurst exponent $H(q)$ of the scale q. (right pane)

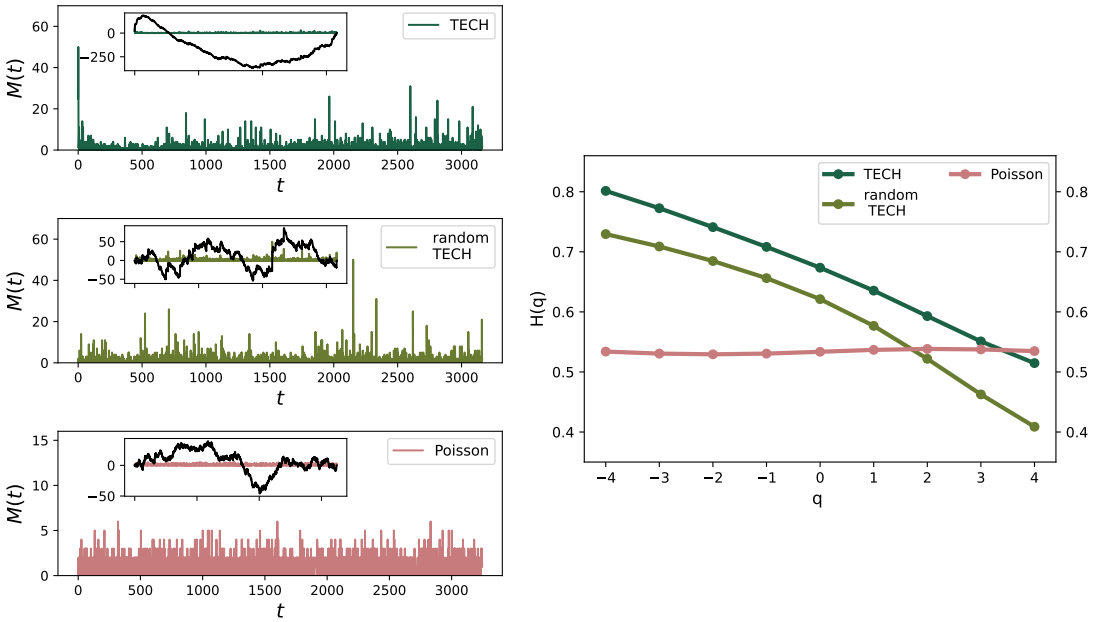

Figure 2.3: TECH signal, the random TECH signal (left pane) and the dependence of multi fractal Hurst exponent $H(q)$ of the scale q. (right pane)

The TECH is a group from the Meetup website that gathers users interested in technology. Using the Meetup website, they organize offline events. The time unit in this time series is an event since then are created links between event. The TECH time series M(t) represents the number of users who joined the TECH community and visited the event for the first time. The time series length is $T = 3162$ steps, and we count $N = 3217$ members in the TECH community for a given period, 2.3. TECH signal has long-range correlations with Hurst exponent $H(q = 2) = 0.6$. Also, we find that TECH is multifractal, as the Hurst exponent is not constant across the scales. The multifractality originates not only from signal trends but also from the broad probability distribution of time series. If we randomize the TECH signal, we find that we can easily destroy trends and cycles, but signal

keeps multifractal properties, meaning that broad probability distribution can not be eliminated. For that reason, we generate the uncorrelated signal from Poissonian probability distribution. The length of this signal is $T = 3246$, while we keep the number of nodes N same as in TECH signal.

2.1.2 D-measure

We can compare the networks with the same number of nodes and links generated with growth signals with different properties. We use growing network model where we vary parameters $-3 < \alpha \leq -1$ and $-3 \leq \beta \leq 1$. We also vary the network density, $L \in \{1, 2, 3\}$. For each set of model parameters α, β, L and each signal $M(t)$, we create the sample of 100 networks. Besides this, for the same set of parameters, we generate the sample of networks with $N = 10000$ and $N = 3217$ nodes grown with constant signal $M(t) = 1$; one node is added to the network at each time step. To examine how different growing signals influence the structure of networks, we use D-measure [16], defined methodology chapter. We equally consider the global and local properties, setting parameter $w = 0.5$. We compare the networks grown with the constant and fluctuating signal with D-measure, for all network pairs between two samples, and averaging the result. The advantage this measure has is that it can measure the distance between two network structures, even if they are generated with the same model; that was not the case with Hamming distance or graph editing distance [16].

Obtained results for D-measure are presented on the figure 2.4. We note that the largest distance between networks is along critical line $\beta(\alpha^*)$ of the aging model. The fluctuations present in the signal mostly influence, the scale-free networks. For networks, away from this line, structural differences exist, but they are much smaller. For gel small world networks, $\beta > \beta^*$, the D-measure is close to zero. Under critical line, $\beta < \beta^*$, D-measure depends on the properties of the signal. If we fix network density L, position of critical line is independent of the properties of the signal. Still, with higher link density, critical line slightly move toward larger β, see figure 2.4.

In the region around the critical line, we find that D-measure depends on the properties of the signal. Multifractal signals TECH has the most considerable impact on network structure; the maximum value of the D-measure is $D_{max} = 0.552$. Similar behaviour is discovered for other multifractal signals, random TECH and MySpace. For networks generated with uncorrelated signals: random MySpace and Poisson, difference exists but it is much smaller.

D-measure rises for lower α. In the case of a constant signal, the number of nodes added to the network is equal for each time step, so at time interval T, the network has MT nodes. In fluctuating signal, the number of nodes added during time interval T vary. In signals, such as TECH, where there are peaks in the number of new users, hubs emerge faster. As we decrease the parameter α, fluctuations in the signal become more critical, and the hubs emerge even for uncorrelated signals. The trends in the real signals further promote the emergence of hubs in the network.

2.1.3 The structure of networks

We examine degree distribution, degree correlations and clustering coefficient of networks generated by real signals. It has been shown that these measures provide a sufficient set for describing the structure of complex networks. Results showed that multifractals influence networks more than monofractals; it is most prominent in scale-free networks.

Figure 2.5 shows properties of networks generated with model parameters $L = 2$, $\alpha = -1.0$, $\beta = 1.5$, that lies on critical line. The degree distributions $P(k)$ of networks generated with real signals TECH and MySpace have super-hubs emerged. Degree distributions generated with randomized and white noise signals do not differ from the degree distribution of networks generated with the constant

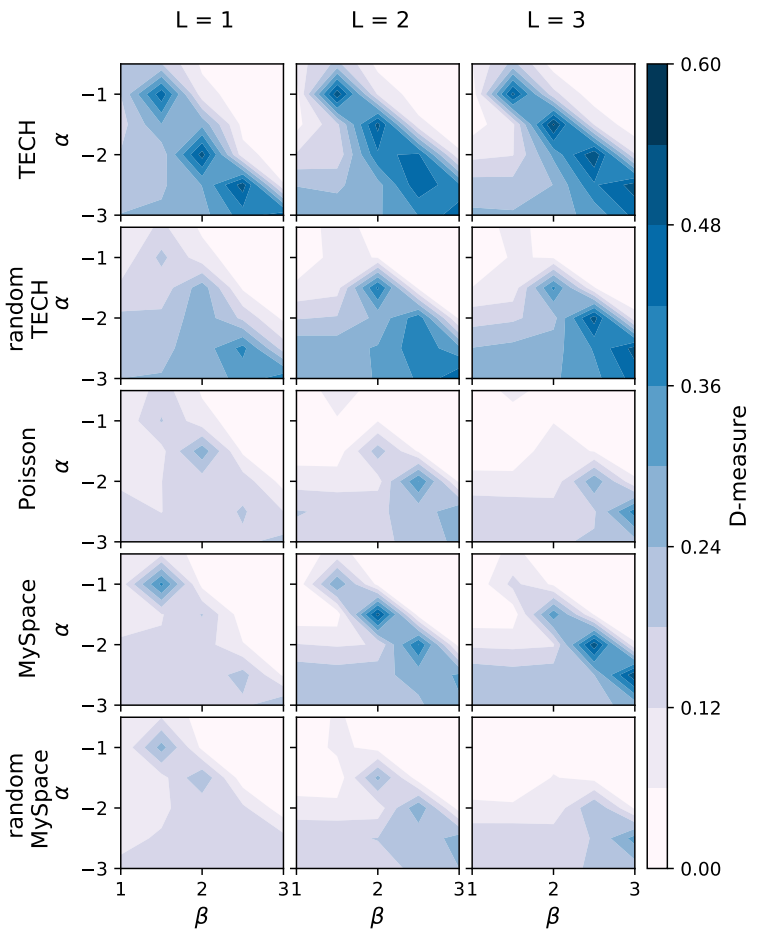

Figure 2.4: The comparison of networks grown with growth signals shown in figures 2.3 and 2.2 versus ones grown with constant signal $M = 1$, for value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in each time step. The compared networks are of the same size.

signal. Networks generated with real signals average neighbouring degree $\langle k \rangle_{nn}(k)$ and clustering coefficient $c(k)$ depend on node degree, while in networks generated with constant and randomized signals, they weakly depend on the degree k.

We also find structural differences between networks, obtained with model parameters under the critical line $\alpha < \alpha^*$, see Figure 2.5. The difference is mostly found for TECH signal. Degree distribution $P(k)$ shows the emergence of hubs in networks grown with TECH signal, while the randomized and Poisson signals are more similar to networks grown with the constant signal. MySpace signal, whose generalized Hurst exponent $H(q)$ weakly depends on scale parameter q and whose long-range correlations and trends are easily destroyed, do not influence the structure of networks more than constant or randomized signal.

The properties of the time-varying signal do not influence the topological properties of small-world gel networks, Figure 2.5. Here model promotes the existence of hubs. As this is the mechanism through which the fluctuations alter the structure of evolving networks, the properties of the signal are not relevant.

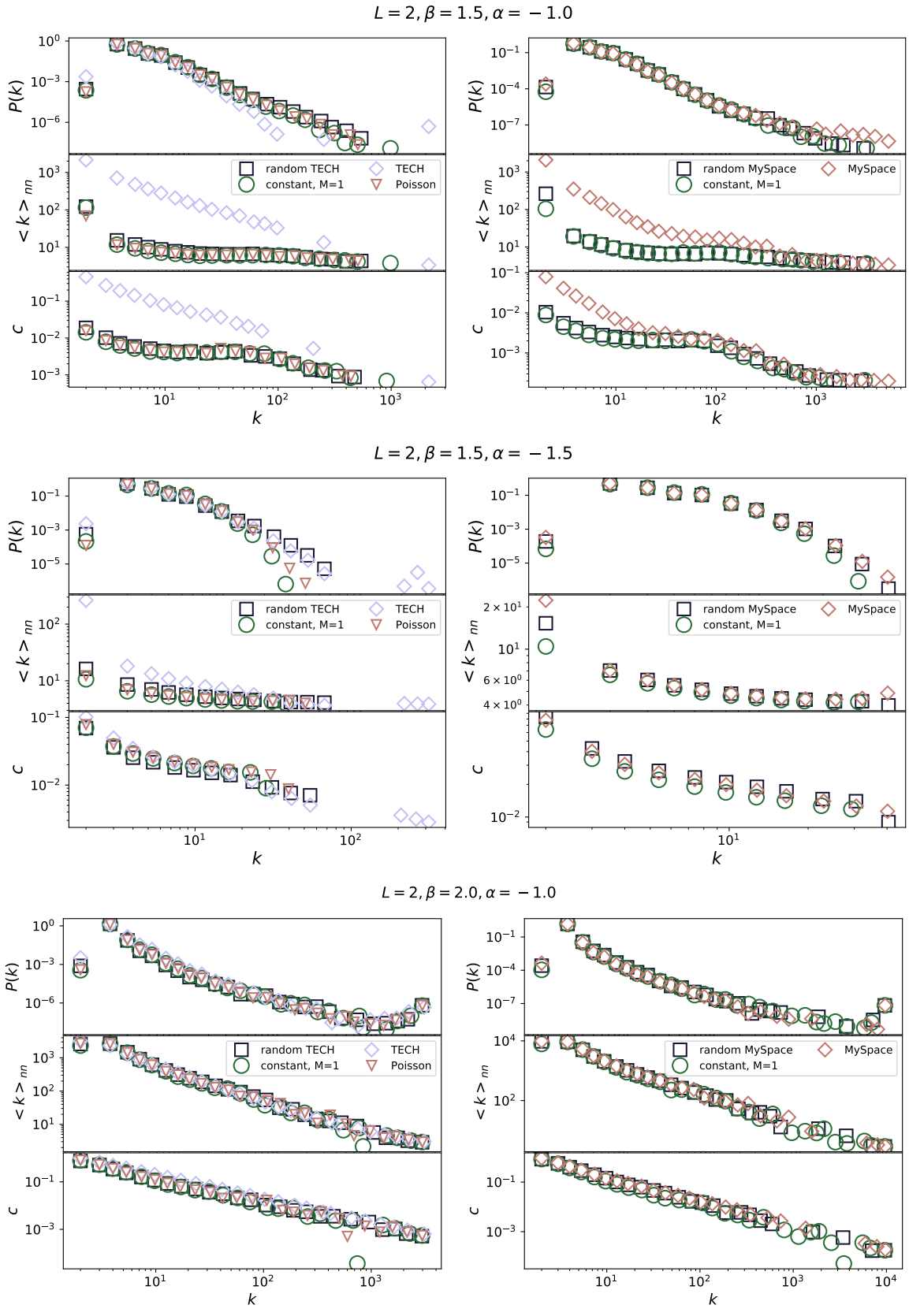

Figure 2.5: Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $\alpha = -1.0$, $\beta = 1.5$ and $L = 2$ for all networks. The networks are from scale-free class. Model parameters have value $L = 2, \alpha = -1.5$, $\beta = 1.5$. The networks have stretched exponential degree distribution. Model parameters have value $L = 2, \alpha = -1.0$, $\beta = 2.0$. Generated networks have small-world properties.

2.2 Long range correlated signals

The previous section showed that the growth signal of real systems has complex dynamics. Besides long-range correlations, we also find multifractal properties, and it is hard to isolate individual effects and analyse their influence separately from each other. When this is the case, synthetic signals with specific characteristics can help to verify our findings in real systems. The long-range correlated properties can be included into time series using Fourier filtering transform method [30].

The long range correlated data have power-law correlations $C(s) =< x_i x_{i+s} >= s^{-\gamma}$ characterized with coefficient γ. Hurst exponent depends on γ as $H = 1 - \frac{\gamma}{2}$. The Fourier transform, gives us the power spectrum of the time series $S(f)$, that is function of the frequency f. For the long-range correlated data it depends on coefficient $\beta = 1 - \gamma$ and has form:

$$S(f) \sim f^{-\beta} \tag{2.2}$$

We can generate the data using Fourier filtering with $\beta = 2H - 1$, as following:

- first generate one-dimensional sequence of uncorrelated random numbers u_i from Gaussian distribution with $\sigma = 1$
- calculate the Fourier transform of the generated sequence, u_q, the spectrum is flat as data correspond to white-noise
- then filter the power spectrum with $f^{-\beta/2}$, so the function will follow power spectrum expected for data with long-range correlations.
- calculate the inverse Fourier transform x_i. It transforms data to the time domain where signal has desired long range correlations

The Fourier filtering method generates the Gaussian distributed data, so data are without broad distributions, nonlinear or multifractal properties. Using this method we generated the signals for differet values of the Hurst exponent, see figure 2.6. The obtained signals are round to integers and mean values of signals are close to 4.

As before, we focus on the region of model phase diagram with negative α and positive β as there is found the transition line from stretched-exponential across scale-free to the small world- gel networks. We take range of parameters $-3 \leq \alpha \leq -0.5$ and $1 \leq \beta \leq 3$ with steps 0.5 and we also vary the the number of links each new node can create $L \in 1, 2, 3$. For each combination of (α, β, L) we generate the sample of 100 networks, and compare the structure of network grown with fluctuating signals with different Hurst exponent $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and constant signal $M = 4$. The results represented by D-measure, shown in the figure 2.7 are obtained averaging the D-measure between all possible pairs of generated networks.

The higher values of D-measure are found in the region of critical line $\beta(\alpha^*)$. The most considerable influence is on networks with scale-free distribution. Comparing D-distance in only one point of phase diagram, for example $L = 1, \alpha = -2.5, \beta = 2.5$, we find that when Hurst exponent is larger, correlations in the signal make bigger impact on the network structure. D-measure between networks grown with signal with Hurst exponent $H = 1.0$ and constant signal is $D(H = 1.0, M = 4) = 0.405$, while between networks grown with signal with $H = 0.8$ and constant signal is $D(H = 0.8, M = 4) = 0.316$. For $\alpha > \alpha^*$ networks have similar structural properties and D-measure is close to 0. In the region of networks with stretched exponential degree distribution $\alpha < \alpha^*$ differences are small.

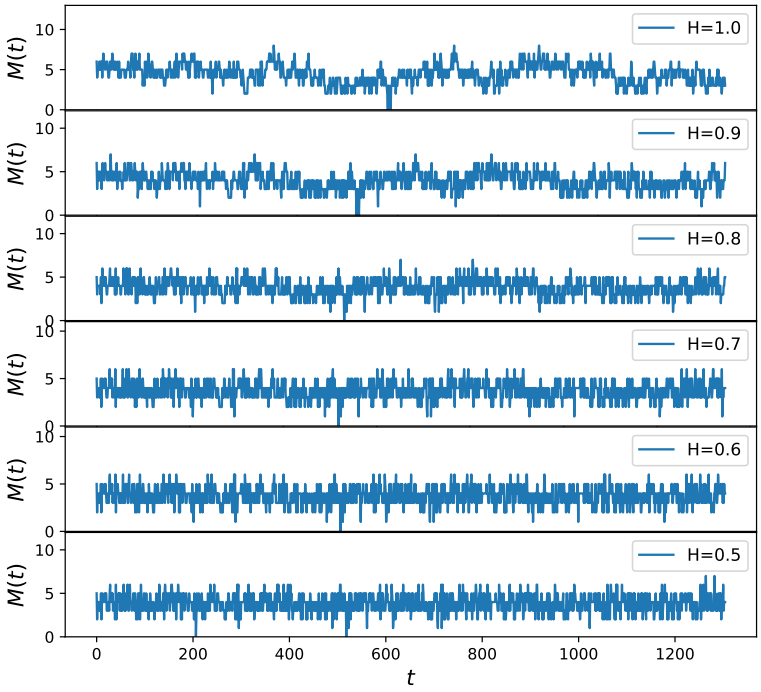

Figure 2.6: Monofractal signals generated with Fourier filtering method for different Hurst exponents

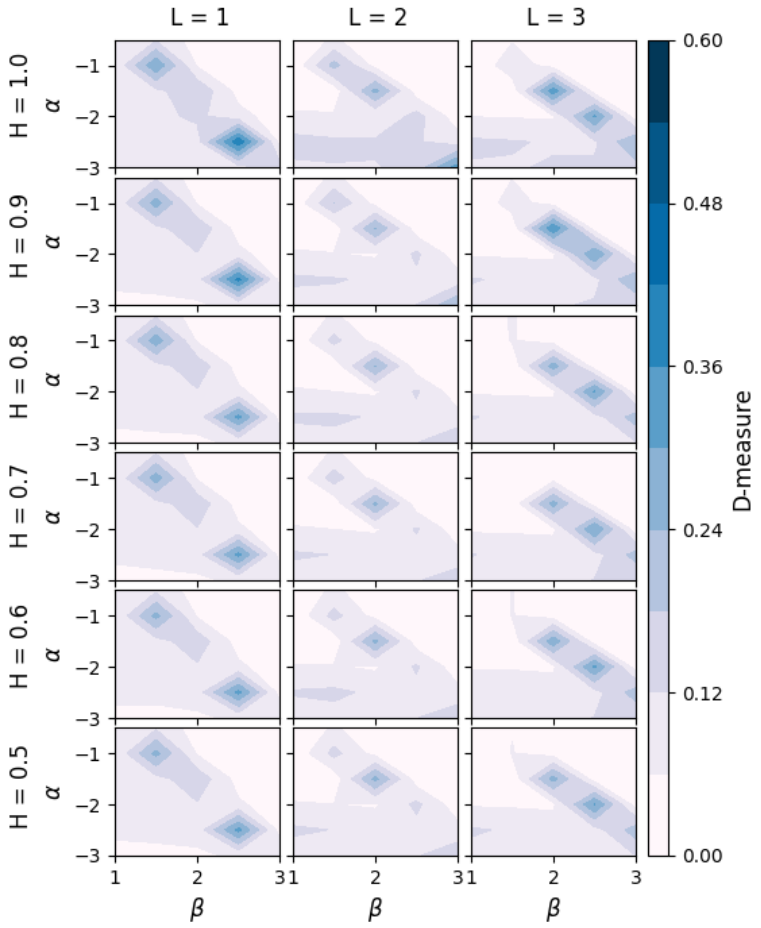

Figure 2.7: D-distance between networks generated with different long-range correlated signals with fixed value of Hurst exponent and networks generated with constant signal M=4.

We further explore the assortativity index and clustering coefficient of generated networks. On figure 2.8 are results for several ageing model parameters that show the difference between networks this model can produce. All networks are disassortative, with a negative degree-degree correlation index. For the values of parameters below critical line, $\alpha = -2.5, \beta = 1.5$ r does not depend on the Hurst exponent. Above the critical line are small-world networks. They are disassortative. The minimum value of the assortativity index is $r = -1$, for $L = 1$, indicating the presence of hubs connecting many nodes. The assortativity index slightly grows with link density.

In the region of critical parameters, the assortativity index depends on the value of the Hurst exponent. The larger influence on the assortativity index have correlated signals, with Hurst exponent $H > 0.8$, so networks become more disassortative, see line for parameters $L = 1, \alpha = -2.5, \beta = 2.5$ in Figure 2.8. The long-range correlations have a stronger effect on the evolution of networks with lower density.

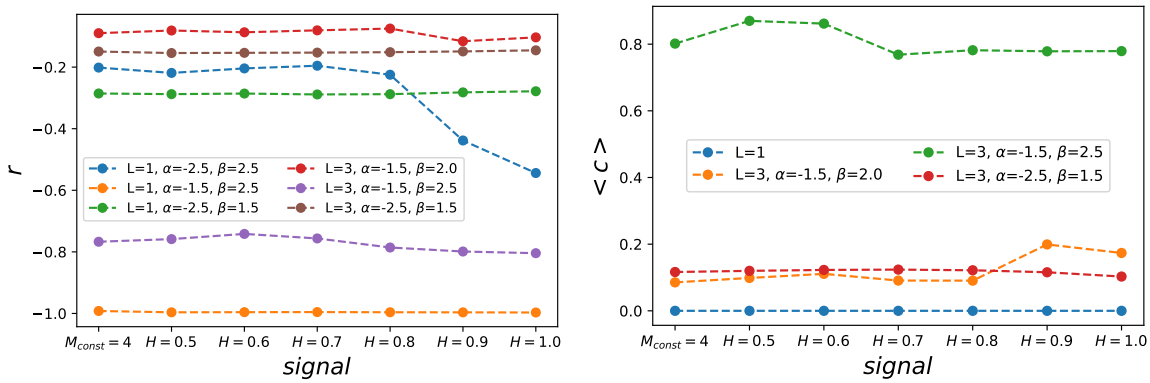

Figure 2.8: Mean assortativity index for networks generated with different with different model parameters α, β, L and different long-range correlated signals with Hurst exponent H.

Figure 2.8 shows the mean clustering coefficient. For $L = 1$, networks are uncorrelated trees, with clustering coefficient 0. For network density $L > 1$, nodes are organized into clusters. Under the critical line, for parameter $L = 3, \alpha = -2.5, \beta = 1.5$, clustering coefficient is constant and low. Similar values are obtained for clustering coefficient for critical parameters $L = 3, \alpha = -1.5, \beta = 2.0$, but for Hurst exponent $H > 0.8$ clustering coefficient increase. Small world networks, $L = 3, \alpha = -1.5, \beta = 2.5$ are clustered, the value of $< c >$ is high. The value of clustering for networks created with the constant signal is 0.8. Networks grown with white noise signal and signal with H=0.6 have higher clustering values, while networks grown with signals with a Hurst exponent larger than 0.6 have the same clustering value, below 0.8.

2.3 Conclusions

Chapter 3

The growth of social groups

3.1 Social groups

Two popular online platforms **Reddit** and **Meetup** are organized into different groups. On Reddit [1], users create subreddits, where they share web content and discussion on specific topics, so their interactions are online through posts and comments. The Meetup groups [2], are also topic-focused, but the primary purpose of these groups is to help users in organizing offline meetings. As meetings happen face-to-face, Meetup groups are geographically localized, so we'll focus on groups created in two towns, London and New York.

The Meetup data cover groups created from 2003, when the Meetup site was founded, until 2018, when using the Meetup API we downloaded data. We extracted the groups from London and New York that were active for at least two months. There were 4673 groups with 831685 members in London and 4752 groups with 1059632 members in New York. For each group, we got information about organized meetings and users who attended them. From there, for each user, we can find the date when the user participated in a group event for the first time; it is considered the date when the user joined a group.

The Reddit data were downloaded from https://pushshift.io/ site. This site collects posts and comments daily; data are publicly available in JSON files for each month. The selected subreddits were created between 2006 and 2011, we also filtered those active in 2017. We removed subreddits active for less than two months. The obtained dataset has 17073 subreddits with 2195677 active members. For each post, we extracted the subreddit-id, user-id and the date when the user created the post. Finally, we selected the date when each user posted on each subreddit for the first time.

3.1.1 The empirical analysis of social groups

For each Meetup group we have information when user attended the group event, while for subreddit we have detailed data about user activity, so we can extract the information when user for the first time created a post. Those dates are considered as timestamp when user joined to group. So both datasets have the same structure: (g, u, t), where t is timestamp when user u joined group g. For each

[1] https://www.reddit.com/
[2] www.meetup.com

time step, we can calculate the number of new members in each group $N_i(t)$, and the group size $S_i(t)$. The group size at time step t is $S_i(t) = \sum_{k=t_0}^{k=t} N_i(t)$, where t_0 is month when group is created. The group size is increasing in time, as we do not have information if the user stopped to be active. Also we calculate the growth rate, as the logarithm of successive sizes $R = log(S_i(t)/S_i(t-1))$.

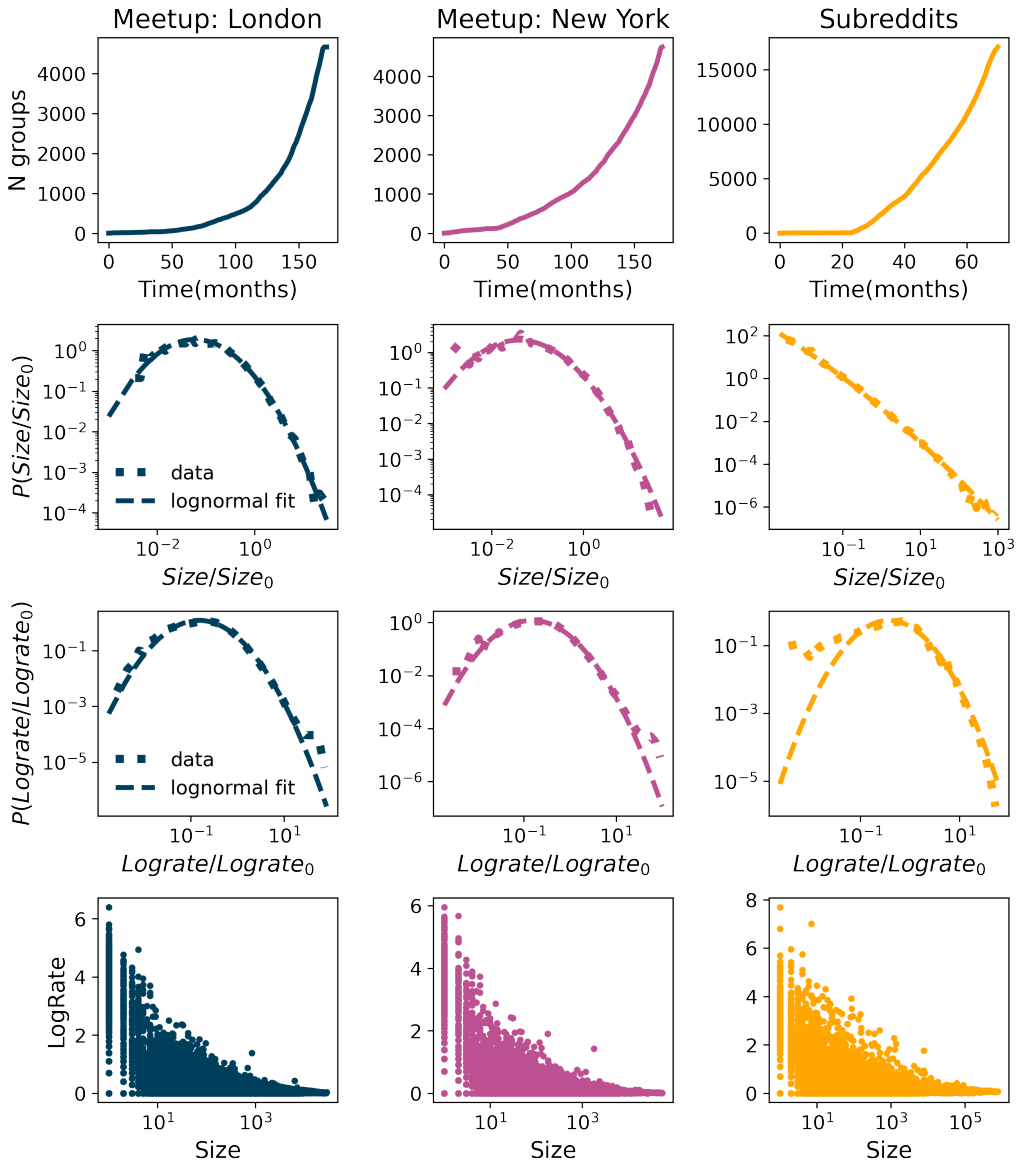

Figure 3.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011 that were active in 2017.

Even though Meetup and Reddit are different online platforms, we find some common properties of these systems; see figure 3.1. The number of groups and the number of new users grow exponentially. Still, subreddits are larger groups than Meetups. The distribution of groups sizes follows the log-normal distribution:

$$P(S) = \frac{1}{S\sigma\sqrt{2\pi}} exp(-\frac{(\ln(S)-\mu)^2}{2\sigma^2}) \tag{3.1}$$

where S is the group size and μ, and σ are parameters of the distribution.

The distributions for Meetup group sizes in London and New York follow similar log-normal distribution, with parameters $\mu = -0.93$, $\sigma = 1.38$ for London and $\mu = -0.99$ and $\sigma = 1.49$ for New

York. The group sizes distribution of Subreddits is a broad log-normal distribution that resembles the power law, it has parameters $\mu = -5.41$ and $\sigma = 3.07$. Still, we used the log-likelihood ratio method and showed that log-normal distribution is better fit for these data then the power-law. In the Result section is given detailed analysis that support this findings.

The simplest model that generates the lognormal distribution is multiplicative process [22]. Gibrat used this model to explain the growth of firms. The main assumption of this model is that growth rates $R = log\frac{S_t}{S_{t-\Delta t}}$ do not depend on the size S and that they are uncorrelated. Further, this imply the lognormal distribution of the sizes, while the distribution of growth rates appears to be normal distribution, [31], [32]. Figure 3.2 shows distribution of the lograte, that follow lognormal distribution, contrary to the Gibrat law. Furthermore, lograte depend on the group size 3.2. For these reasons the Gibrat law can not explain the growth of online social groups [33, 34].

The growth of online social groups has universal behavior. It is independent on the size of the group. If we aggregate the groups created in the same year y, and each group size normalize with average size $< S^y >$, $s_i^y = S_i^y / < S^y >$ we will find that group sizes distributions for the same dataset and different years fall on the same line, figure 3.2. The same characteristics are observed for the distribution of the normalized lograte 3.2. The growth is universal in time, and the group sizes distribution do not change from year to year.

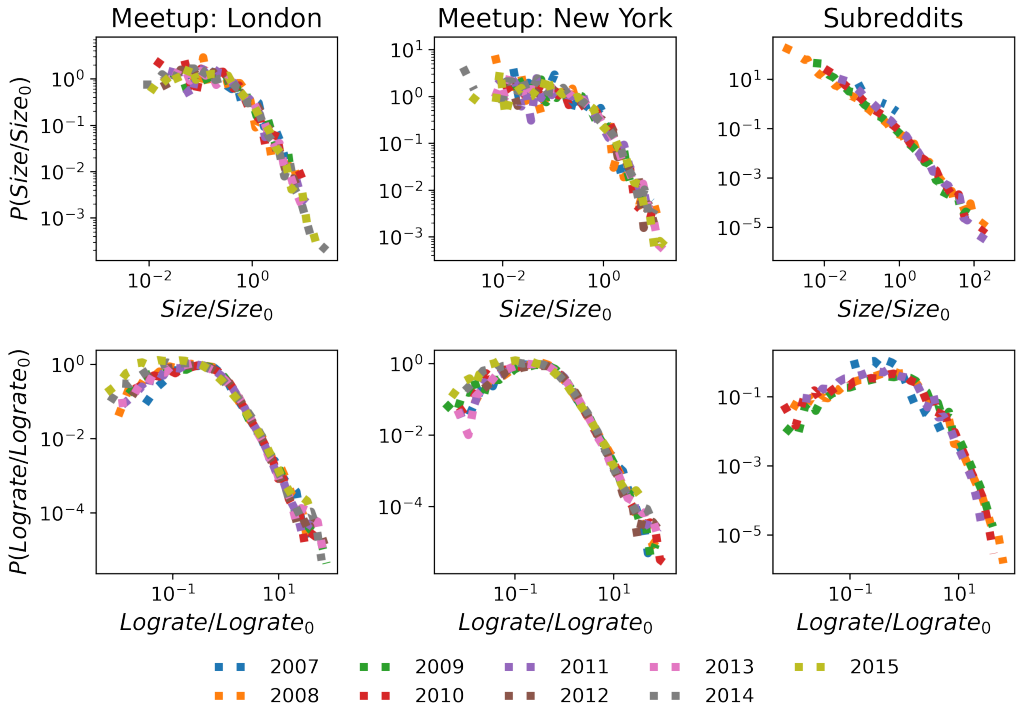

Figure 3.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

3.2 The model

Meetup and Reddit engage members in different activities. Still, there are some underlying processes same in both systems. Each member can create new groups and join existing ones. Both systems grow in the number of groups and users, and each user can belong to arbitrary number of groups. In

the previous section, we identified the universal patterns in the growth of social groups, but it appears that the growth can not be modelled with Gibrat law.

The complex network models allow us to simulate the growth of these systems considering all types of members' activities. We can identify how model parameters shape the growth process by varying linking rules. Regarding the user's group choice, it was shown that social connections play an important role [35, 36]. On the other hand, users can be driven by personal interests. Diffusion between groups could also be enhanced with rich-get-richer phenomena, where users tend to join larger groups. With a complex network model, we can easily incorporate the nonlinear growth in the number of users and groups, as it is an important parameter that shapes the structure and dynamics of the complex network [37, 38, 39].

The evolution of the social groups has been studied using the co-evolution model in the reference [36]. This model consists of two evolving networks: the bipartite network, which stores connections between users and groups and the affiliation network of social connections. At each time step, active users create new connections in the affiliation network; i.e. they make new friends. They also join existing groups or create new ones, which updates the bipartite network. The group selection can be random with probability proportional to the group size; otherwise, the group is selected through social contacts. Using this model, authors have reproduced the power-law group size distribution found in several communities, such as Flickr or LiveJournal. The empirical analysis of Meetup and Reddit groups showed that group size distribution could be log-normal, meaning that some different mechanisms control the growth of the groups.

We propose a model that is based on the co-evolution model. The main difference between those two models is how model parameters are defined. First of all, in the co-evolution model user becomes inactive after period t_a, which is drawn from an exponential distribution with the rate λ, while in our model probability that the user is active is constant, and the same for each user. The second difference is how groups are chosen. While in the co-evolution model probability that the user selects a group through social linking depends on the friend's degree, we give preference to groups where a user has a larger number of social contacts. We also modified the rules for random linking, so users choose a group with uniform probability.

3.2.1 Groups growth model

The representation of the model is given in figure 3.3. The model consists of two networks:

- bipartite network $\mathcal{B}(V_U, V_G, E_{UG})$, where V_U is set of users, V_G set of groups and E_{UG} set of links between users and groups, where link $e(u, g)$ indicates that user u is member of group g.
- social network $\mathcal{G}(V_U, E_{UU})$ describes the social connections $e(u, v)$ between users u and v, and $V(U)$ is set of users same as in bipartite network.

The bipartite and social networks evolve. At each step, new users $N_U(t)$ are added to the network. It is how the set of users V_U in the bipartite and social network can grow. At arrival, each new member connects to a randomly selected user in the social network G. This allows new members to choose a group based on social contacts [35]. The activity of old members is a stochastic process; old members are activated with probability p_a. The set of active users $\mathcal{A}_U$ has new members $N_U(t)$ and old members who decided to be active in that time step.

The active users can create a new group with probability p_g. By this, group node g is added to the set of group nodes V_G in bipartite network B. If an active user does not create a new group, it will join the existing one with probability $1 - p_g$, see lower panel on figure 3.3. When the user creates a new group or joins an existing one, the link $e(u, g)$ is made in the bipartite network B.

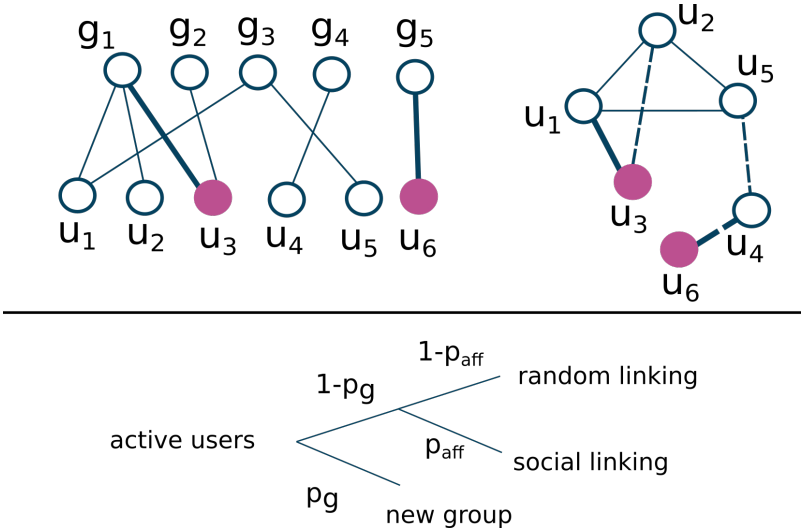

Figure 3.3: The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model schema, where p_g is probability that user create new group, while p_{aff} is probability that group choice depends on the social connections. **Example:** member u_6 is a new member. First it will make random link with node u_4, and then with probability p_g makes new group g_5. With probability p_a member u_3 is active, while others stay inactive for this time step. Member u_3 will with probability $1 - p_g$ choose to join one of old groups and with probability p_{aff} linking is chosen to be social. As its friend u_2 is member of group g_1, member u_3 will also join group g_1. Joining group g_1, member u_3 will make more social connections, in this case it is member u_1.

When joining existing groups, users may be influenced by social connections. This linking happens with probability p_{aff}. The second case is that the user chooses a random group with probability $1 - p_{aff}$.

Social linking depends on the properties of a bipartite and social network. The networks can be represented with matrices B and A, so if a link between two nodes exists, they have element 1. The neighbourhood of user u, $\mathcal{N}_u$ in a bipartite network is a set of groups in which the user is a member. Similarly, we define the neighbourhood of group g as N_g, as a set of users who belong to the group. From there, we can define the probability P_{ug} that the user u will choose group g. This probability is proportional to the number of social contacts that the user has in the group.

$$P_{ug} = \sum_{u_1 \in \mathcal{N}_g} A_{uu_1} \tag{3.2}$$

After selecting group g, user u is introduced to new members in the group and can make new social contacts. In the simplest case, we could assume that all members belonging to a group are connected. However, previous research on this subject [28, 40, 36] has shown that the existing social connections of members in a social group are only a subset of all possible connections. We select X random members u_i from group g and make new connections in social network $e(u, u_i)$.

The model parameters p_a and p_g are important for controlling the number of users and groups. With larger parameter values p_a, more users become active, and the number of links in bipartite and social networks grows faster. Parameter p_g controls the rate at which new groups are created. For example, if $p_g = 0$, users will not create new groups. Also, if $p_g = 1$, users will only create new groups, and the resulting network will consist of star-like subgraphs. In real systems we do not expect

extreme values for probabilities p_a and p_g. First, not all members are constantly active, and we do not find a burst in the creation of the groups. From real data, we notice that there is always a higher number of users than groups in social systems. The parameter p_{aff} how users choose groups, and with higher p_{aff} social connections become more important.

3.2.2 Dependence of the group size distribution on model parameters

Before applying the group growth model on Meetup and Reddit, we consider the system where at each time step, a constant number of users is added $N(t) = 30$. We also fix the probability that user is active to $p_a = 0.1$, so we can in more details explore the influence of parameters p_g and p_{aff}. We plot the group size distribution after 60 steps of simulation. The values of p_g and the p_a influence the number of groups, their maximum size, and the shape of group size distribution. With probability $p_g = 0.1$, users create large number of groups, over 10^4, while with $p_g = 0.5$ they are on the order of magnitude 10^5.

Figure 3.4 show the obtained group size distributions with power-law and log-normal fits. For lower value of parameter $p_g = 0.1$ and $p_{aff} = 0$, users join randomly chosen groups. Group size distributions are approximated with log-normal. When the affiliation parameter is larger, $p_{aff} = 0.5$, the log-normal distribution becomes broader, and so on, we find the larger maximum group size.

If we increase the parameter $p_g = 0.5$, every second active user will create a group. At this group creation rate, the group size distribution deviates from log-normal, but it is not explained with power-law either, right column on figure 3.4.

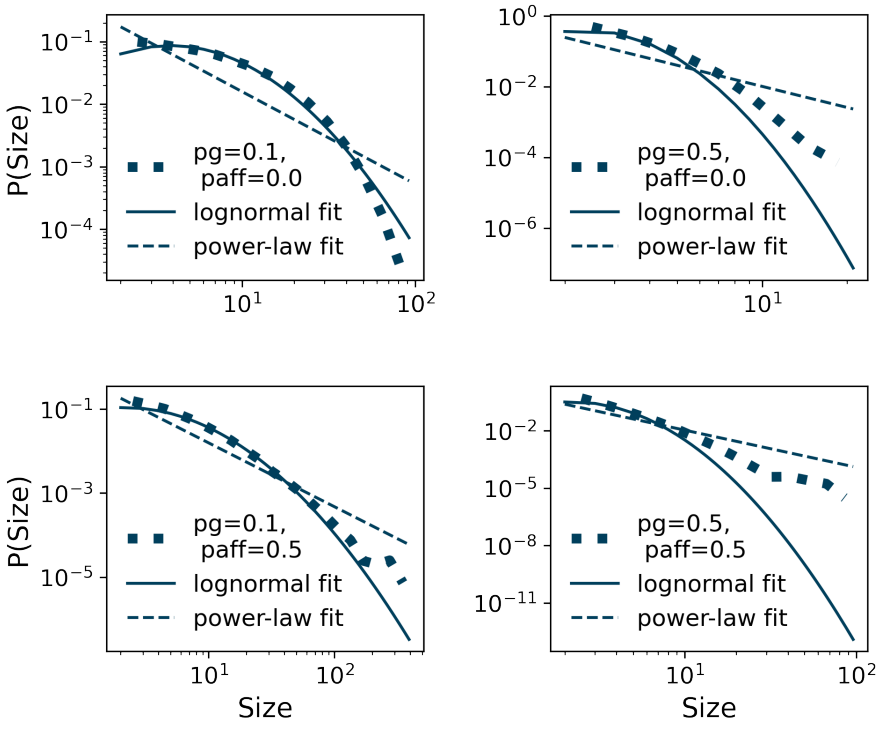

Figure 3.4: The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

Finally, we compare how group size distribution depends on different rules in random linking. In our model, the probability that the user chooses a random group is uniform. In contrast, in the co-evolution model [36], probability depends on the group size, as in the preferential attachment model.

Instead of random linking, if we incorporate preferential linking, so users with probability $1 - p_{aff}$ tend to choose larger groups, group size distribution changes significantly. Similar to the co-evolution model, we find the power-law distribution. Figure 3.5 shows the results from a model where we add a constant number of new users at each time step. The probabilities p_a and p_g are fixed, and affiliation parameter takes values 0, 0.5 and 0.8. If we consider random linking , top panel on figure 3.5, the distribution becomes broader with larger p_{aff}. On the other hand, with preferential linking, group size distribution is a power-law and the p_{aff} parameter does not have a large impact on the distribution shape.

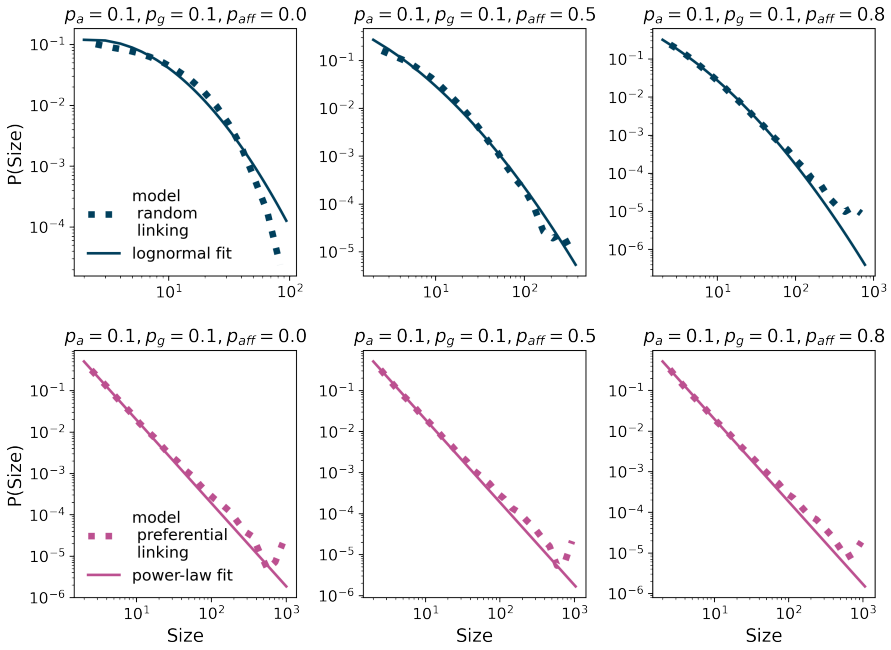

Figure 3.5: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff}. With probability, $1 - p_{aff}$, users choose a group randomly. The group sizes distribution (top row) is described with a log-normal distribution. With higher affiliation parameter, p_{aff}, distribution has larger width. The bottom row presents the case where with probability $1 - p_{aff}$ users have a preference toward larger groups. For all values of parameter p_{aff}, we find the power-law group sizes distribution.

3.3 Results

The social systems do not grow at constant rate. In Ref. [39] authors have shown that features of growth signal influence the structure of social networks. For these reasons we use the real growth signal from Meetup groups located in London and New York, and Reddit community to simulate the growth of the social groups in these systems. Figure 3.6 top panel shows the time series of the number of new members that join each of the three systems each month. All three systems have relatively low growth at the beginning, and than the growth accelerates as the system becomes more popular.

We also use empirical data to estimate p_a, p_g and p_{aff}. Probabilities that old members are active p_a and that new groups are created p_g can be approximated directly from the data. Activity parameter p_a is the ratio between the number of old members that were active in month t and the total number of members in the system at time t. Figure 3.6 middle row shows the variation of parameter p_a during the considered time interval for each system. The values of this parameter fluctuates between 0 and

0.2 for London and New York based Meetup groups, while its value is between 0 and 0.15 for Reddit. To simplify our simulations we assume that p_a is constant in time, and estimate its value as its median value during the 170 months for Meetup systems, and 80 months of Reddit system. For Meetup groups based in London and New York $p_a = 0.05$, while Reddit members are more active on average and $p_a = 0.11$ for this system.

Figure 3.6 bottom row shows the evolution of parameter p_g for the three considered systems. The p_g in month t is estimated as the ratio between the groups created in month t $Ng_{new}(t)$ and the total number of groups that month $Ng_{new}(t) + Ng_{old}(t)$, i.e., $p_g(t) = \frac{Ng_{new}(t)}{N_{new}(t)+N_{old}(t)}$. We see from Fig. 3.6 that $p_g(t)$ has relatively high values at the beginning of the system's existence. This is not surprising. At the beginning these systems have relatively small number of groups and often cannot meet the needs for content of all their members. As the time passes, the number of groups grows, as well as content offerings within the system, and members no longer have a high need to create new groups. Figure 3.6 shows that p_g fluctuates less after the first few months, and thus we again assume that p_g is constant in time and set its value to median value during 170 months for Meetup and 80 months for Reddit. For all three systems p_g has the value of 0.003.

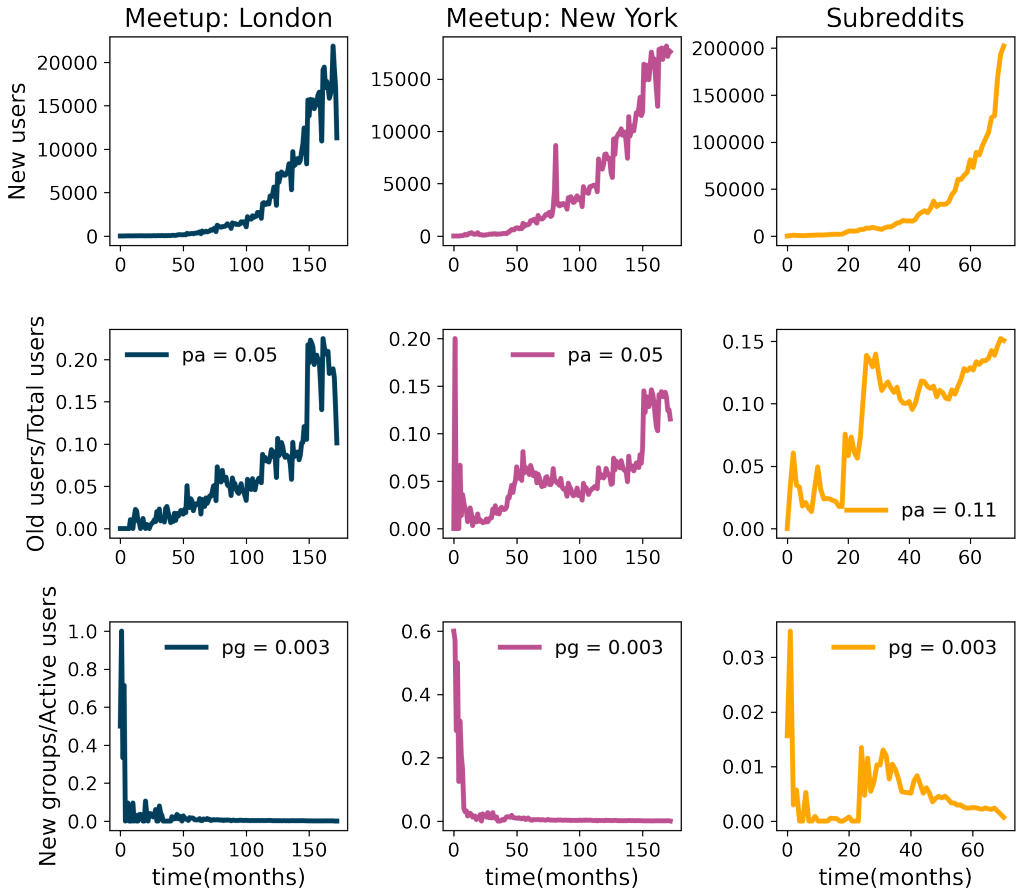

Figure 3.6: The time series of number of new members (top panel), ratio between old members and total members in the system (middle panel), and ratio between new groups and active members(bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

The affiliation parameter p_{aff} is not possible to estimate directly from the empirical data. For these reasons, we simulate the growth of social groups each of the three systems with the time series of new members obtained from the real data and estimated values of parameters p_a and p_g, while we vary the value of p_{aff}. For each of the three systems, we compare the distribution of group sizes obtained from simulations for different values of p_{aff} with ones obtained from empirical analysis

using Jensen Shannon (JS) divergence. The JS divergence [41] between two distributions P and Q is defined as

$$JS(P,Q) = H\left(\frac{P+Q}{2}\right) - \frac{1}{2}\left(H(P) + H(Q)\right) \tag{3.3}$$

where $H(p)$ is Shannon entropy $H(p) = \sum_x p(x)log(p(x)$. The JS divergence is symmetric and if P is identical to Q, $JS = 0$. The smaller the value of JS divergence, the better is the match between empirical and simulated group size distributions. The Table 3.1 shows the value of JS divergence for all three systems. We see that for London based Meetup groups the affiliation parameter is $p_{aff} = 0.5$, for New York groups $p_{aff} = 0.4$, while the affiliation parameter for Reddit $p_{aff} = 0.8$. Our results show that social diffusion is important in all three systems. However, Meetup members are more likely to join groups at random, while for the Reddit members their social connections are more important when it comes to choice of the subreddit.

p_{aff}	JS cityLondon	JS cityNY	JS reddit2012
0.1	0.0161	0.0097	0.00241
0.2	0.0101	0.0053	0.00205
0.3	0.0055	0.0026	0.00159
0.4	0.0027	**0.0013**	0.00104
0.5	**0.0016**	0.0015	0.00074
0.6	0.0031	0.0035	0.00048
0.7	0.0085	0.0081	0.00039
0.8	0.0214	0.0167	**0.00034**
0.9	0.0499	0.0331	0.00047

Table 3.1: Jensen Shannon divergence between group sizes distributions from model (in model we vary affiliation parameter paff) and data.

Figure 3.7 shows the comparison between the empirical and simulation distribution of group sizes for three considered systems. We see that empirical distributions for Meetup groups based in London and New York are perfectly reproduced by the model and chosen values of parameters. In the case of Reddit, the distribution is very broad, and the tail of distribution is well reproduced by the model. The bottom row of Fig. 3.7 shows the distribution of logarithmic values of growth rates of groups obtained from empirical and simulated data. We see that the tails of empirical distributions for all three systems are well emulated by the ones obtained from the model. However, there are deviations which are the most likely consequence of using median values of parameters p_a, p_g, and p_{aff}.

3.3.1 Distributions fit

We compute the log-likelihood ratio R, and p-value between different distributions and log-normal fit [42] to determine the best fit for the group size distributions. Distribution with a higher likelihood is a better fit. The log-likelihood ratio R then has a positive or negative value, indicating which distribution represents a better fit. To choose between two distributions, we need to calculate p-value, to be sure that R is sufficiently positive or negative and that it is not the result of chance fluctuation from the result that is close to zero. If the p-value is small, $p < 0.1$, it is unlikely that the sign of R is the chance of fluctuations, and it is an accurate indicator of which model fits better.

Table 3.2 summarizes the findings for empirical data on group size distributions from Meetup groups in London, Meetup groups in New York and Reddit. Using the maximum likelihood method, we obtain the parameters of the distributions [43]. The results indicate that log-normal distribution

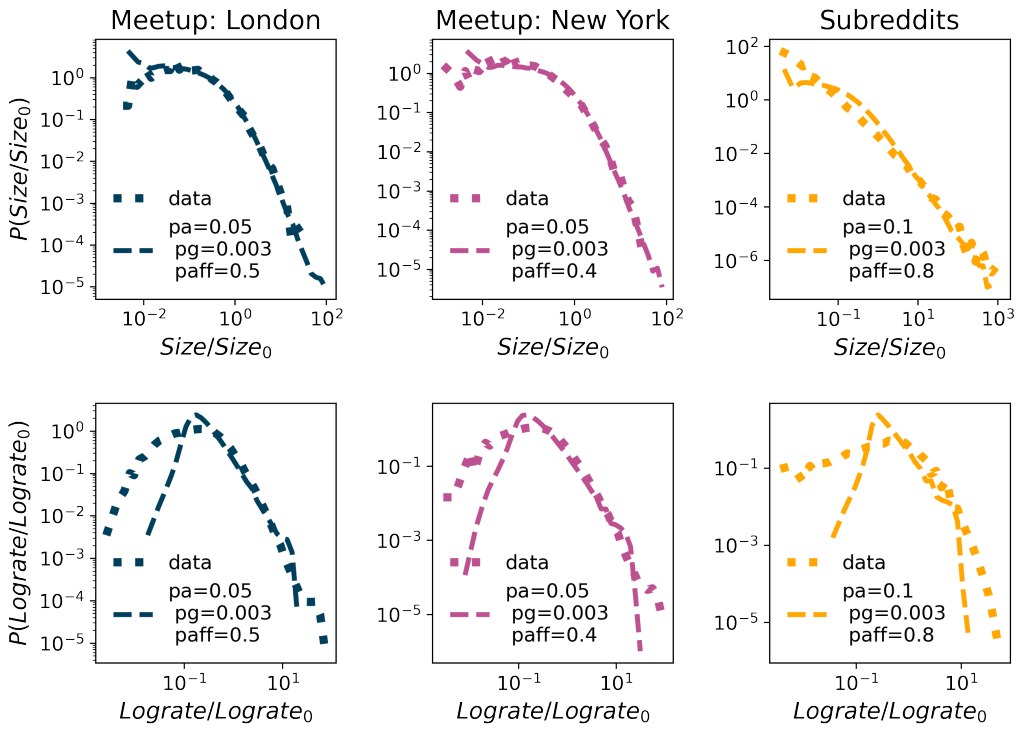

Figure 3.7: The comparison between empirical and simulation distribution for group sizes (top panel) and lograte (bottom panel).

is the best fit for all three systems. Figure 3.8 shows the distributions of empirical data as well as log-normal fit on data. For Meetup data, we present fit on stretched exponential distribution, which very well fits a large portion of data. For subreddits, distribution is broad and, potentially, resembles power-law. Still, log-normal distribution is a more suitable fit.

Table 3.2: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **groups sizes** of Meetup groups in London, New York and in Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

distribution	Meetup city London		Meetup city NY		Reddit	
	R	p	R	p	R	p
exponential	-8.64e2	8.11e-32	-8.22e2	6.63e-26	-3.85e4	1.54e-100
stretched exponential	-3.01e2	1.00e-30	-1.47e2	7.78e-8	-7.97e1	5.94e-30
power law	-4.88e3	0.00	-4.57e3	0.00	-9.39e2	4.48e-149
truncated power law	-2.39e3	0.00	-2.09e3	0.00	-5.51e2	2.42e-56

We use the same methods to estimate the fit for simulated group size distributions on Meetup groups in London, New York, and Subreddits. Table 3.3 shows the results of the log-likelihood ratio R and p-value between different distributions. We conclude that log-normal distribution is most suitable for simulated group size distributions. Plotting log-normal and stretched exponential fit on data, Fig. 3.9 we confirm our observations.

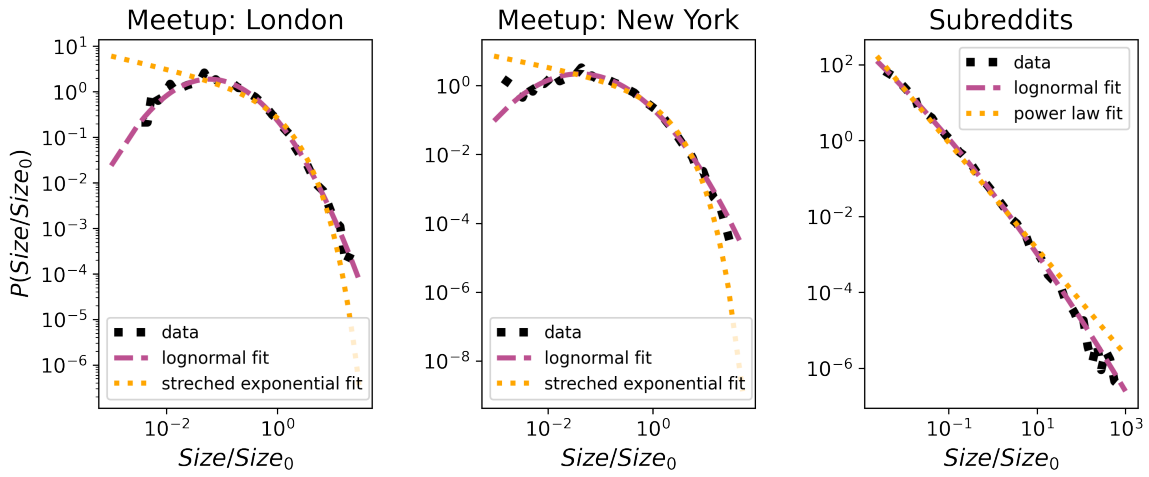

Figure 3.8: The comparison between log-normal and stretched exponential fit to London and NY data, and between log-normal and power law for Subreddits. The parameters for log-normal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits $\mu = -5.41$ and $\sigma = 3.07$.

Table 3.3: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **simulated group sizes** of Meetup groups in London, New York and Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

distribution	Meetup city London		Meetup city NY		Reddit	
	R	p	R	p	R	p
exponential	-6.27e4	0.00	-5.11e4	0.00	-1.26e5	7.31e-125
stretched exponential	-1.01e4	1.96e-287	-6.69e3	1.46e-93	-1.39e4	0.00
power law	-2.29e5	0.00	-3.73e5	0.00	-4.38e4	0.00
truncated power law	-9.28e4	0.00	-1.55e5	0.00	-9.12e4	0.00

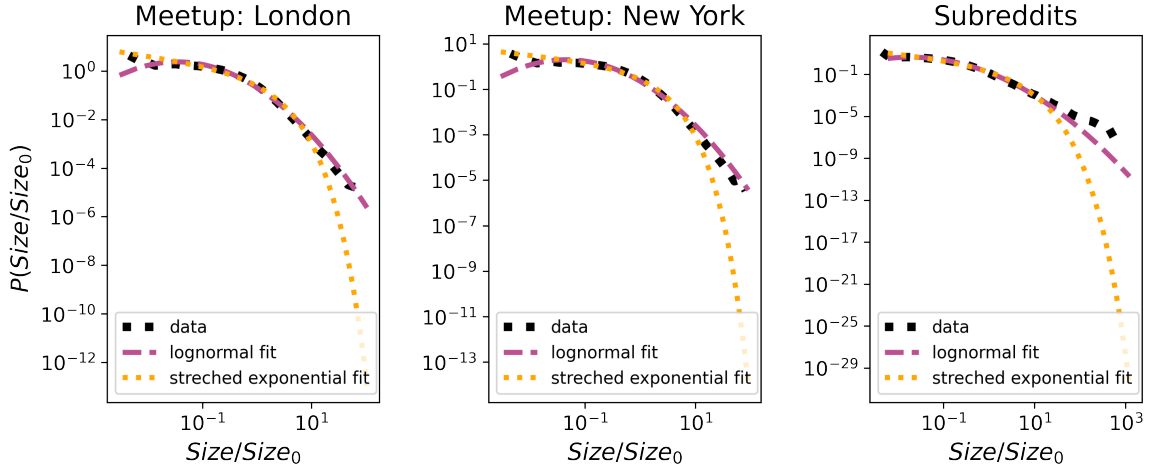

Figure 3.9: The comparison between lognormal and stretched exponential fit to simulated group sizes distributions. The parameters for log-normal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma = 1.53$.

Chapter 4

The role of trust in knowledge based communities

Information and communications technologies (ICTs) have enabled faster and easier creation and sharing of knowledge. Furthermore, they have provided access to a large amount of data which enabled a detailed study of their emergence and evolution [38], as well as user's roles [44], patterns of their activity [45, 46, 47]. However, relatively small attention was given to sustainability of SE communities. Most of the research was focused on the activity and factors that influence the increase of the users' activity in these communities. Factors such as need for experts and the quality of their contributions have been thoroughly investigated [48]. It was shown that growth of communities and mechanisms that drive it may depend on the topic around which the community was created [49].

The **Stack Exchange** is a network of question-answer websites on diverse topics. In the beginning, the focus was on computer programming questions with StackOverflow [1] community. Its popularity led to the creation of the Stack Exchange network that these days counts more than 100 communities on different topics. The SE communities are self-moderating, and the questions and answers can be voted, allowing users to earn Stack Exchange reputation and privileges on the site.

The new site topics are proposed through site Area51 [2], and if the community finds them relevant, they are created. Every proposed StackExchange site needs interested users to commit to the community and contribute by posting questions, answers and comments. After a successful private beta phase site reaches the public beta phase, other members are allowed to join the community. The site can be in the public beta phase for a long time until it meets specific SE evaluation criteria for graduation. Otherwise, it may be closed with a decline in users' activity.

We focused analysis on four pairs of SE communities with the same topic. Astronomy, Literature and Economics are active communities [3] The first time, these communities were unsuccessful and thus closed. We also compare closed Theoretical Physics with the Physics site, considering that those two topics engage similar type of users.

[1] More information about StackOverlflow is available at: https://stackoverflow.co/ and broad introduction to StackExchange network is available at: https://stackexchange.com/tour.

[2] Visit https://area51.stackexchange.com/faq for more details about closed and beta SE communities and the review process.

[3] Astronomy, Literature and Economics graduated on December 2021 and during our research, they were still in the public beta phase.

4.1 Network properties of Stack Exchange data

On Stack Exchange sites, the interaction between users happens through posts. As we are interested in examining the characteristics of the users, we map interaction data to the networks. Using complex network theory, we can quantify the properties of obtained networks and compare different SE communities, e.g. alive and closed SE sites.

In the user interaction network, the link between two nodes, user i and j, exists if user i answers or comments on the question posted by user j, or user i comments on the answer posted by user j. The created network is undirected and unweighted, meaning that we do not consider multiply interactions between users or the direction of the interaction.

First approach is to aggregate all interactions in the first 180 days, and study the properties of static network. Many local and global network measures are dependent [1], and it was shown that degree distribution, degree-degree correlations and clustering coefficient are sufficient for description of the properties of complex networks [50].

We calculate the **degree distribution**, figure 4.1, and compare the distributions of active and closed communities of the same topic. Degree distributions between active and closed communities follow similar lines.

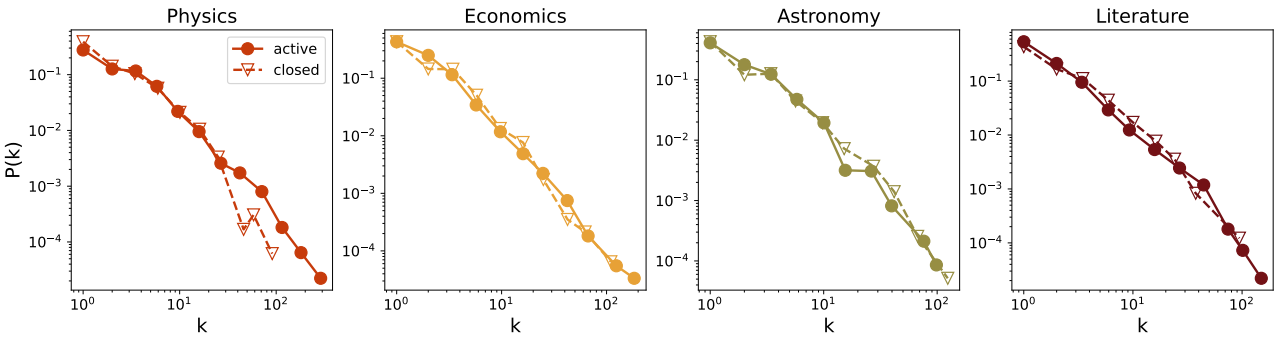

Figure 4.1: Degree distribution.

If we take look into **neighbor degree** dependece on the node degree $k_{nn}(k)$, figure 4.2 we find that there are structural differences between networks formed in the active and closed communities. On average k-degree users in active communities have neighbors with larger degree than it is case in closed communities. The results are consistent for Physics, Economics and Literature. For Astronomy we find different behavior, where the $k_{nn}(k)$ distributions of closed communities are on the top of distributions of the active one.

The **clustering coefficient** of a node quantifies the average connectivity of between its neighbours and cohesion of its neighborhood [1]. It is a probability that two neighbours of a node are also neighbours, and is calculated using the following formula:

$$c_i = \frac{e_i}{\frac{1}{2}k_i(k_1 - 1)} \,. \tag{4.1}$$

Here e_i is the number of links between neighbours of the node i in a network, while $\frac{1}{2}k_i(k_i - 1)$ is the maximal possible number of links determined by the node degree k_i. The clustering coefficient of network C is the value of clustering averaged over all nodes. Study on dynamics of social group growth shows that that links between one's friends that are members of a social group increase the probability that that individual will join the social group [40]. Furthermore, successful social diffusion

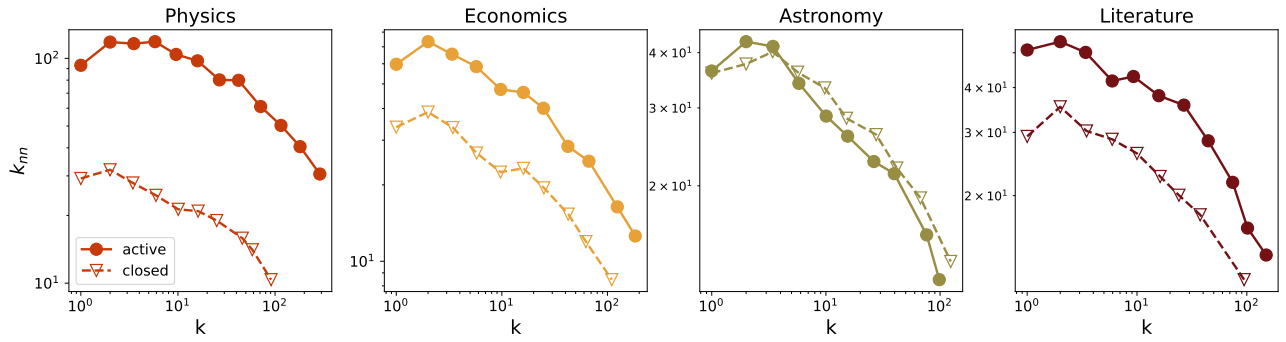

Figure 4.2: Neighbour degree.

typically occur in networks with high value of clustering coefficient [51]. These results suggest that high local cohesion should be a characteristic of sustainable communities. The dependence of the clustering coefficient on the node degree is shown on figure 4.3. As expected we find that active communities are more clustered.

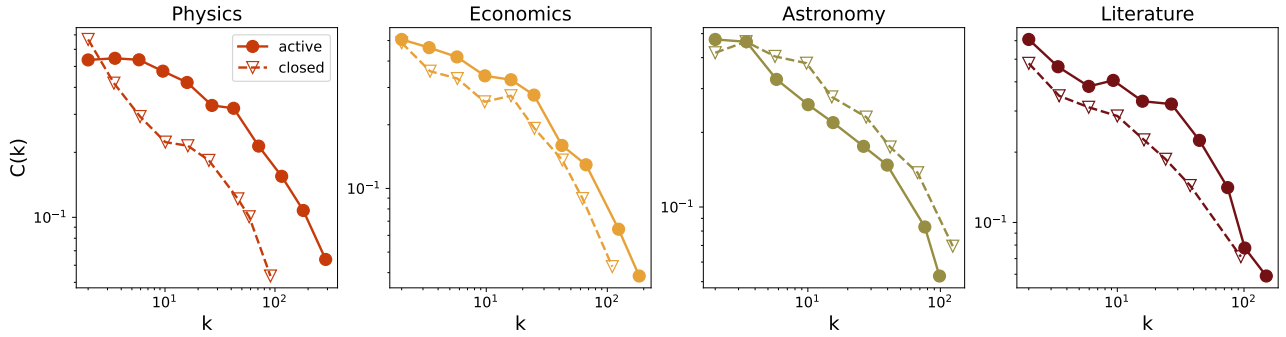

Figure 4.3: Clustering coefficient.

Instead of creating a static network from the data in the first 180 days of community life, we study how network snapshots evolve. At each time step t, we create network snapshot $G(t, t+\tau)$, for time window of the length τ. We fix the time window to $\tau = 30$ days and slide it by $t = 1$ day through time. Discussion of how the length of the sliding window influences the results is given in appendix A. Sliding the time window by one day, we can capture changes in the network structure daily, as two 30 days consecutive networks overlap significantly.

Here we investigate how clustering coefficient in a SE community is changing with time by calculating its value for all network snapshots. We compare the behavior of clustering for active and closed communities on the same topic in order to better understand how cohesion of these communities is changing over time. Figure 4.4 shows the evolution of mean clustering coefficient for all eight communities. All communities that are still alive are clustered, with the value of mean clustering coefficient higher than 0.1. Physics, the only launched community, has the value of clustering coefficient above 0.2 for the first 180 days.

During larger part of the observed period, the clustering coefficient of an active community is higher compared to the clustering coefficient of its closed pair. If we compare active communities with their closed counterpart, the closed communities have higher value of the mean clustering coefficient in the early phase while later communities that are still active have higher values of clustering coefficient. These results suggest that all communities have relatively high local cohesiveness, and

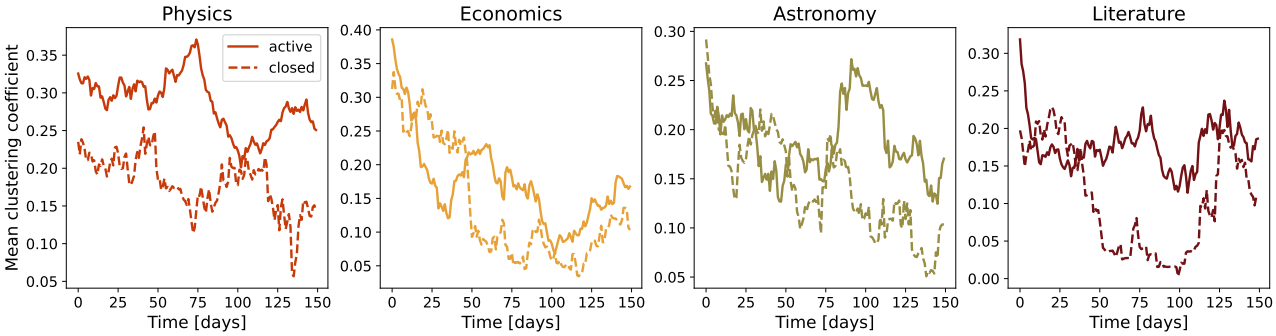

Figure 4.4: Mean clustering coefficient.

that lower values of clustering coefficient in the later phase of community life may be an indicator of its decline.

4.2 Core-periphery structure

Previous research on Stack Exchange communities have attempted to explain how different types of users interact. In Question-Answer communities are expected popular and casual users [45, 49]. Popular users generate the majority of interactions in the system, they are experts in community and take care on answering questions and engage the discussions through comments. As popular users they considered the 10% of the most active users, and showed that popular users are highly connected not only among themselves but also with casual users.

We tested this theory on all eight communities. We focused on 30 days sub-networks and showed how the number of links per node among popular users and between popular and casual users, evolve over time, figure 4.5. We also compare active and closed communities of the same topic, so links per nodes in active sites are larger than in closed communities.

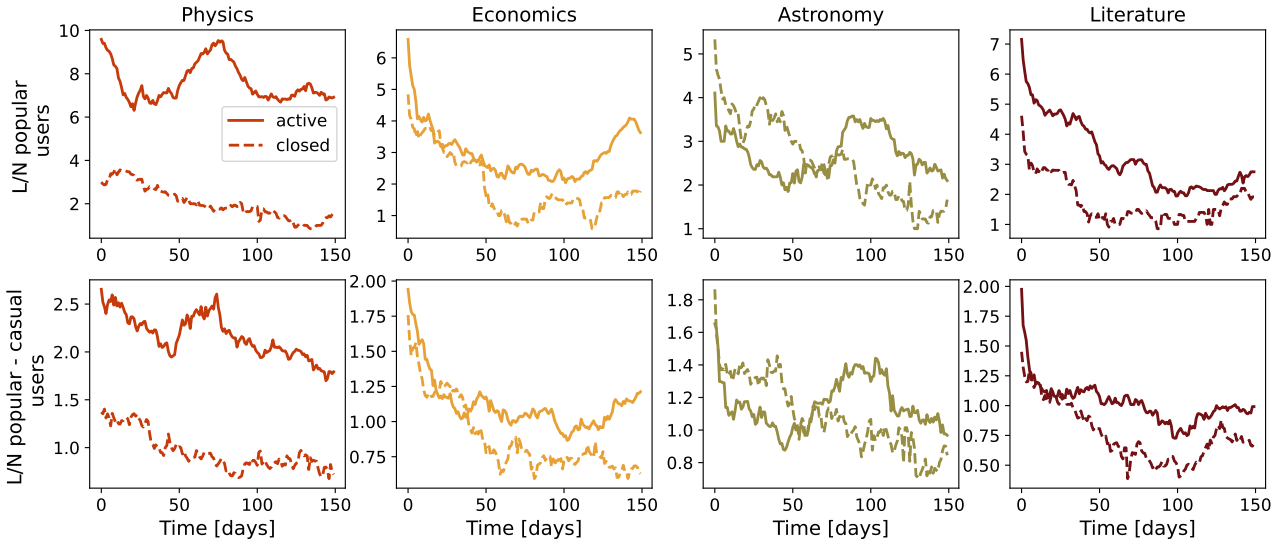

Figure 4.5: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users).

Although, we find the difference between active and closed communities, the split according to

10% most active users does not guaranty that all popular users will be considered. Furthermore, the smaller group of frequently active users is similar to the core users in core-periphery structure. This is why we are going to detect the core of the each 30day network. By this, separation is based on the network structure, and is more consistent, as using algorithmic approach we optimize the connectivity inside the core, periphery and among them. Core-periphery structure has core that is densely connected group of nodes, while the periphery has low density [52, 19].

We use Stochastic Block Model (SBM) to infer the core-periphery structure of each 30 days network snapshot and analyses how core structure evolve over time. The SBM algorithm is adapted for inferring the core-periphery structure, [19]. For each 30 days network we run the sample of 50 iterations and choose the model parameters according to minimum description length. As stochastic models start from the random configuration, they can converge to different states, so we analyzed the stability of the inferred structures. More details are given in the appendix. We found that obtained structures differ, but minimum description length does not fluctuate too much. Also, different similarity measures between inferred core configurations take values higher than 0.9, indicating that core structure is stable.

Number of users in core of active communities is higher than in closed communities, top panel on figure 4.6. On the other hand we do not find strong difference between the fraction of core users in the closed and active communities. Furthermore, the fraction of users in core differ from the 10%, and it is constantly changing, bottom panel 4.6.

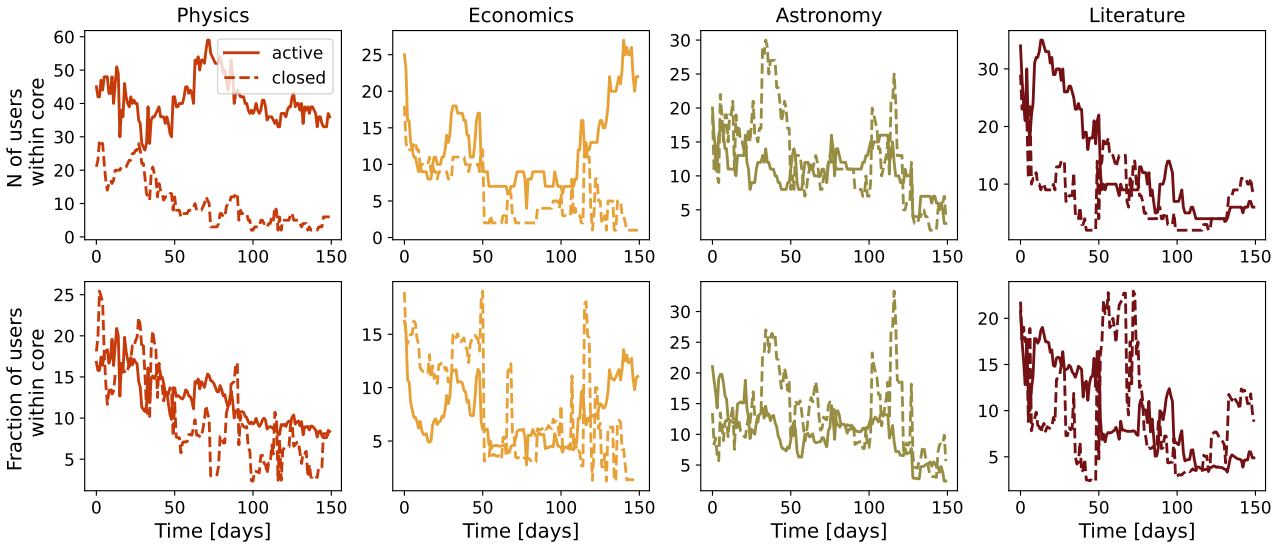

Figure 4.6: Just for reference size of the core (top) and fraction of users in core (bottom). Solid lines - active sites; dashed lines - closed sites.

The number of users is constantly changing. To quantify, the stability of the core structure we compute the Jaccard's coefficient between core users in networks at time points t_1 and t_2. The Jaccard coefficient range from 0 to 1, so the larger values of Jaccard index indicate the more similar cores. The highest values are found around diagonal elements where we compare networks closer in time, see figure 4.7. The core membership is changing over time, and it is more frequent in the closed communities.

The average Jaccard index between cores in networks separated by time interval $t_i - t_j$ with the standard deviation confidence interval are shown in figure 4.8. The Jaccard index decreases with relative time difference between networks faster in closed communities. The relatively high overlap between distant networks confirms that active networks have more stable core.

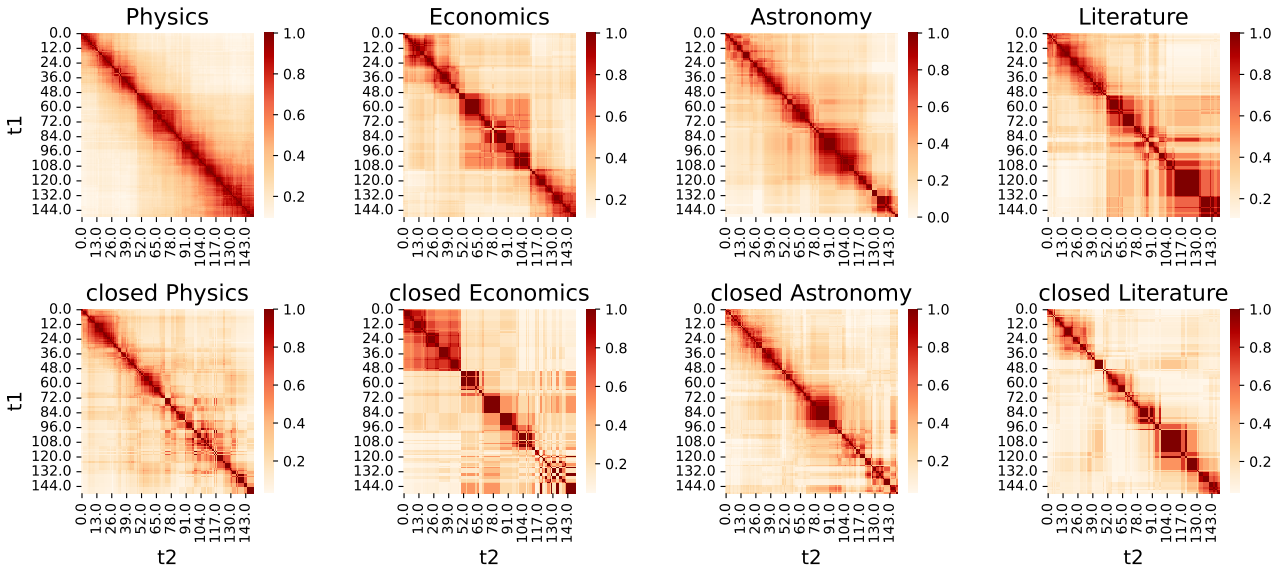

Figure 4.7: Jaccard index between core users in sub-networks at time points $t1$ and $t2$

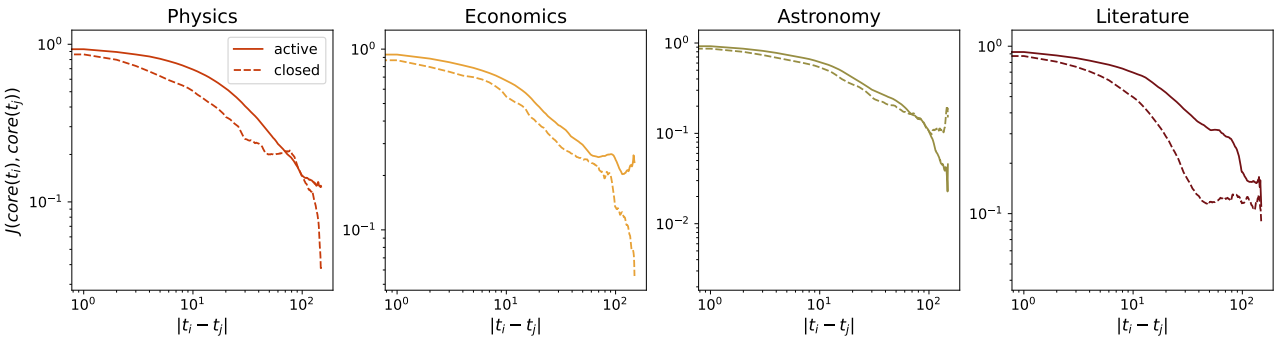

Figure 4.8: Jaccard index between core users in 30days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$

Finally, we examine how the connectivity of the users in the core and between core and periphery evolve over time. On figure we show the L/N in the core, that is proportional to the average degree of the network $2L/N$. The Physics community has more than twice larger connectivity than closed Theoretical Physics. For Literature we also find higher connectivity, but at the end of observation period the values become, the connectivity in active site drops and becomes similar as in closed one. For Economics and Astronomy the difference between active and closed site is not so clear. At the beginning of the period for the sites on economic topic, connectivity is similar, After 50 days of community life, connectivity in active communities is starting to rise, while in the case of closed economics it is dropping. For Astronomy, the connectivity is higher in closed communities, in the first 50 days, After this, period we find the sudden rise in the connectivity of active astronomy, but again it is dropping and becomes comparable to the connectivity values in closed site. The similar conclusions can be drawn for the connectivity between core and periphery. The largest difference between active and closed site is observed for Physics. When it comes to active communities that are still in the beta phase, they either have the same core-periphery connectivity as their closed counter part, or as in the case of Astronomy, their periphery is weaker connected to the core during the first 50 days of their life, see Fig. 4.9.

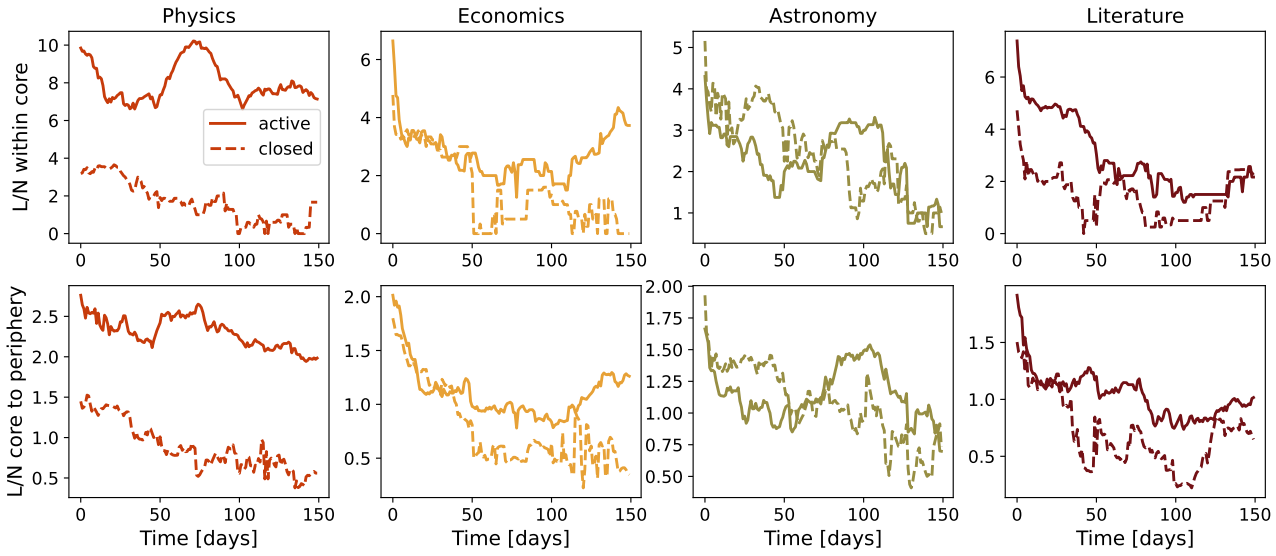

Figure 4.9: Links per node in core and links per node between core and periphery.

4.3 Dynamical Reputation model

We further explore the difference between active and closed communities through the dynamical reputation model. With this model we calculate the reputation of each user in the community. The reputation is directly connected with the collective trust in the network, that is important property of sustainable communities. From network properties we found that active communities are more cohesive. The core of active communities is also more stable.

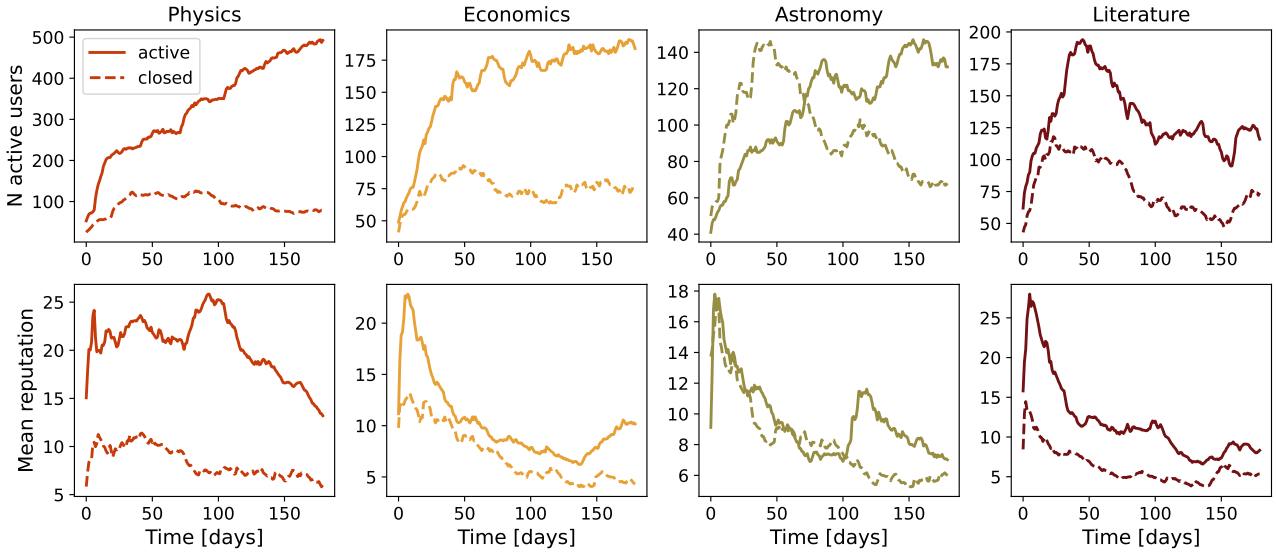

Figure 4.10: Dynamic Reputation on the four pairs of Stack Exchange websites: Astronomy, Literature, Economics, Physics and Theoretical Physics.

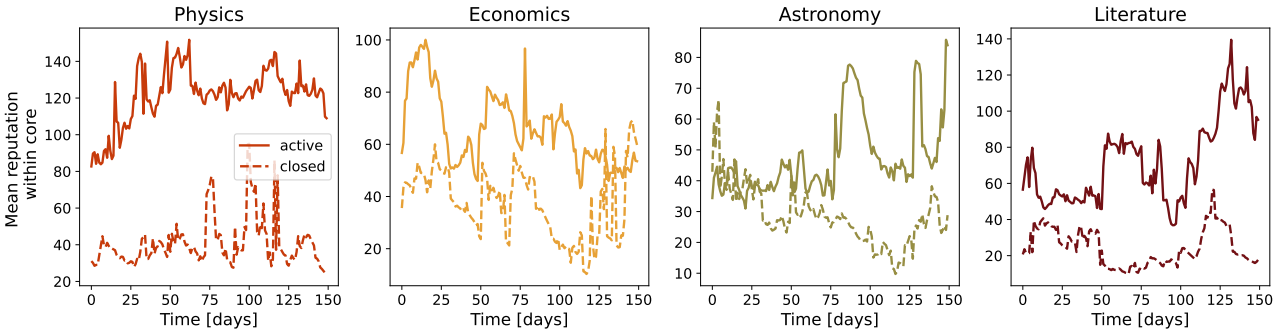

Figure 4.11: Dynamical reputation within core.

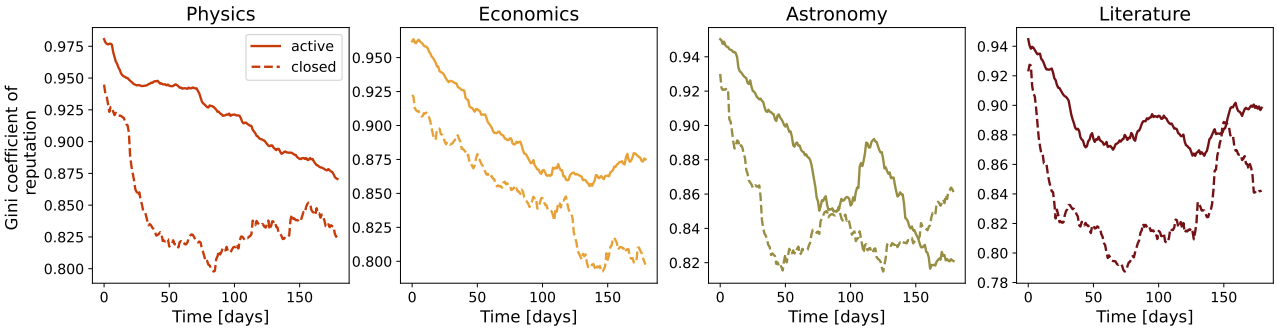

Figure 4.12: Gini index of dynamic reputation within population

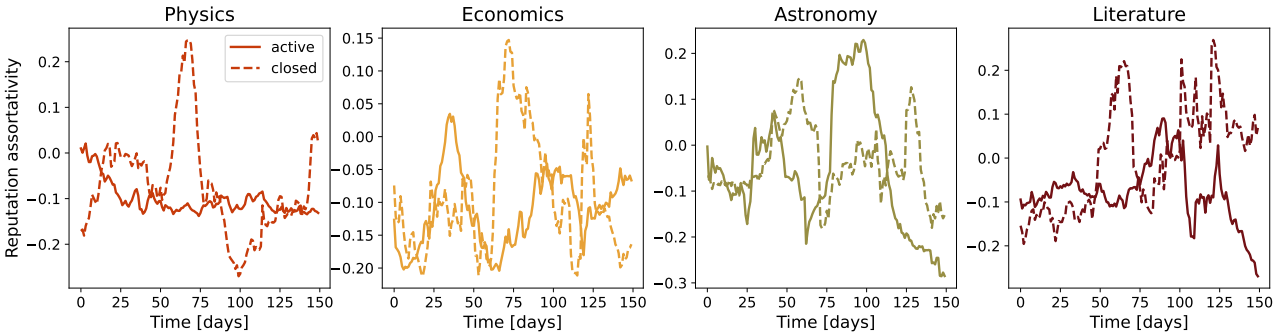

Figure 4.13: Dynamic Reputation assortativity in the network of interactions (questions, answers, comments, unweighted, undirected network). Solid lines - active sites; dashed lines - closed sites.

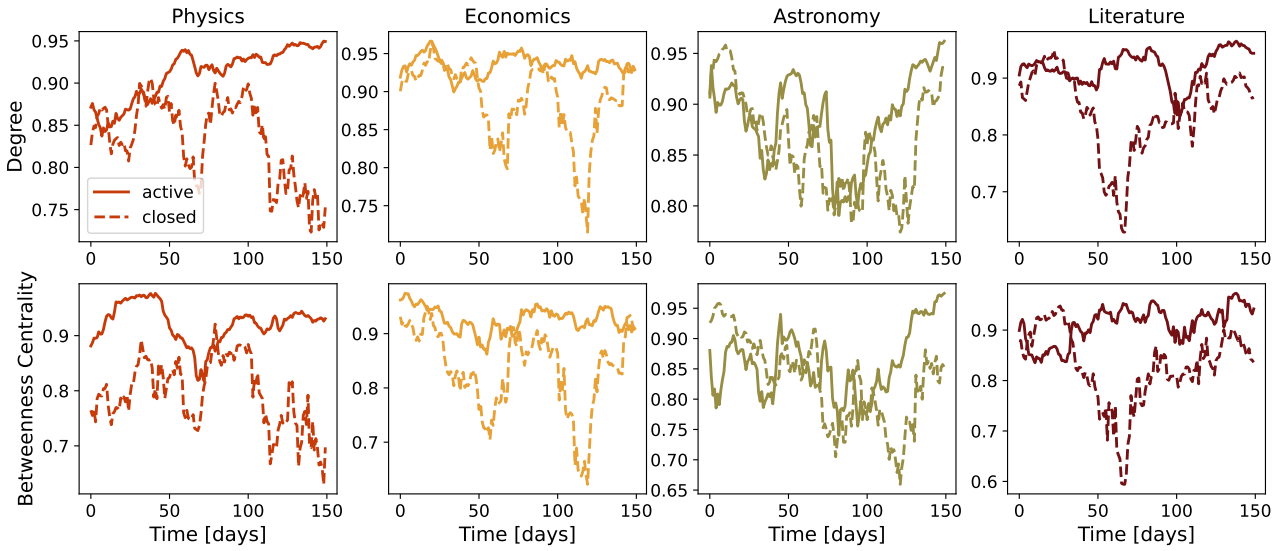

Figure 4.14: Coefficient of correlation between users' Dynamic Reputation and users' network degree (top) and users's betweenness centrality (bottom). Solid lines - active sites; dashed lines - closed sites.

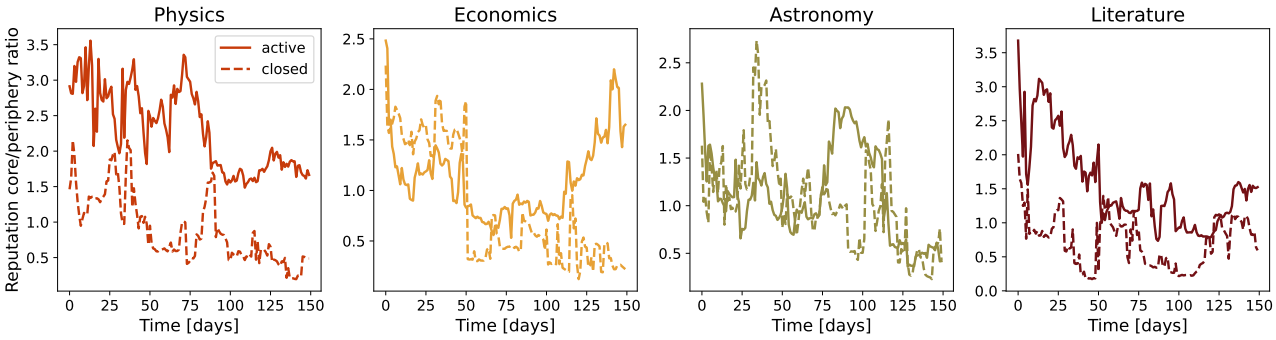

Figure 4.15: Ratio between the total reputation within network core and periphery. Solid lines beta communities, dashed lines area 51 communities.

Appendix A

Stack Exchange

Stack Exchange data are public and regularly released. As closed communities were active between 180 and 210 days, we extracted only first 180 days of data. Given that first few months can be crucial for further development of the community [53], we are interested in early evolution of Stack Exchange sites.

Detailed information about questions, answers, and comments are available for each SE community. Each post is labelled with a unique ID, the user's ID who made the post, and creation time. On Stack Exchange, users interact on several layers: Those interactions are considered positive.

- posting an answer on the question; for every question, we extract IDs of its answers
- posting a comment on the question or answer; for every question and answer, we selected IDs of its comments
- accepting answer; for each question, we selected the accepted answer ID

Even though posts can be voted and downvoted, information about a user who voted is absent, so we do not consider these interactions between users. Comments can not be downvoted, while we find only around 3% negatively voted answers and questions, Table A.1.

Table A.1: Percentage of negatively voted interactions

Site	Status	Questions	Answers
Physics	Beta	5%	4%
	Closed	1%	2%
Astronomy	Beta	3%	3%
	Closed	2%	1%
Economics	Beta	4%	4%
	Closed	7%	4%
Literature	Beta	2%	5%
	Closed	2%	1%
Average		3.2%	3%

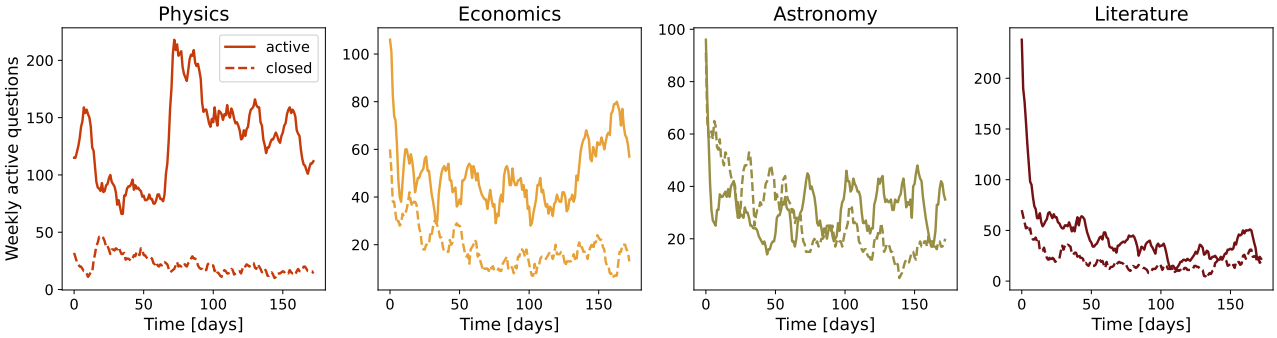

Figure A.1: Number of active questions within 7 days sliding windows. Solid line - active sites; dashed lines - closed sites.

A.1 Comparison between active and closed SE communities

Table A.2 compares the first 180 days between closed and active communities. When it comes to basic statistics, active communities had larger number of users, questions, answers and comments. Another simple indicator if community is going to graduate or decline can be time series of active questions for period of 7 days in Figure A.1. The question is active if had at least one activity, posted answer ot comment during previous seven days. We find that live communities have larger number of active questions after first three months. Still, this difference is smaller for literature and astronomy. For astronomy we observe that closed community had higher number of active questions in the early period of community life.

Table A.2: Community overview for first 180 days, Number of users n_u, number of questions n_q, number of answers n_a, number of comments n_c

Site	Status	First Date	n_u	n_q	n_a	n_c
Astronomy	Closed	09/22/10	336	474	953	1444
	Beta	09/24/13	405	644	959	2170
Economics	Closed	10/11/10	275	368	458	1253
	Beta	11/18/14	648	1024	1410	3553
Literature	Closed	02/10/10	284	318	523	1097
	Beta	01/18/17	478	910	907	3301
Physics	Closed	09/14/11	281	349	564	2213
	Launched	08/24/10	1176	2124	4802	15403

Similarly, the official Stack Exchange community evaluation process considers simple metrics [1]. To determine the the success of sites they measure how many questions are answered, how many questions are posted per day, and how many answers are posted per question. There are two measures: the number of avid users and the number of visits that are not easily interpreted from the data. The site is *healthy* if it has 10 questions per day, 2.5 answers per question and more than 90% of answered

[1] https://stackoverflow.blog/2011/07/27/does-this-site-have-a-chance-of-succeeding/

questions. For less than 80% of answered questions, 5 questions per day and 1 question per answer site *needs some work*.

We calculated Stack Exchange statistics for astronomy, economics, literature and physics and results are presented in the Table A.3. After observed period of 180 days only live physics is healthy site while other live communities are at least in two criteria labeled as *okay*. Closed sites mostly *need some work*, the exception is closed astronomy. For example it has *excellent* percent of answered questions and *okay* answer ratio.

Table A.3: Community overview for first 180 days according to SE criteria

Site	Status	Answered	Questions per day	Answer ratio
Astronomy	Closed	**95** %	2.62	2.02
	Beta	**96** %	3.57	1.49
Economics	Closed	68 %	2.04	1.25
	Beta	84 %	5.66	1.37
Literature	Closed	79 %	1.77	1.65
	Beta	74 %	5.04	1.10
Physics	Closed	83 %	1.93	1.64
	Beta	**93** %	**11.76**	**2.74**
Stack Exchange criteria	excellent	> 90 %	>10	> 2.5
	needs some work	< 80 %	< 5	< 1

This simple measurements presented in tables A.2 and A.3 and in figure A.1 do not provide us clear indications about community sustainability. Only for physics topic the difference between active and closed community is evident, while for other communities it is not so clear. Thus, we need deeper insights into structure and dynamics of these communities to understand. The structure of social interactions within communities and dynamics of collective trust may provide better explanation why some communities succeed and other died.

Appendix B

Choice of model parameters

B.1 Selection of dynamical reputation model parameters

One of the largest drawbacks of DIBRM is the parameter tuning problem. In previous applications of the model [26, 54] there was no single best set of parameter values for modeling dynamic reputation in Stack Exchange communities. For example, in [54] the best approximation of the official Stack Exchange reputation is obtained with $t_a = 2, \beta = 1, \alpha = 1.4$ which means there is no active forgetting factor. In our application of DIBRM to SE communities we opted for a different set of parameter values. Details of parameter search and tuning are presented in SI.

For basic reputation contribution of a single interaction we selected $I_{bn} = 1$ and at the same time this is the threshold value of an active user. This value is intuitive as every interaction has initial contribution of +1 to user's reputation, although the previous works have used values of +2 and +4. Following the previous work and after examining the median/average time between subsequent interactions of the same user, we selected $t_a = 1$, which also means that reputation in our model will be updated every day during the time-window of the analysis, regardless of whether the user is active or not. To emphasize the bursts of activity and frequent recent interactions, cumulative factor has a larger value $\alpha = 2$. Finally, the most delicate parameter is the forgetting factor, which at the same time determines the weight of past interactions and the reputational punishment due to user inactivity. Here we need to select the value of parameter β so we include the forgetting due to inactivity but not to penalize is too much. In Fig. A1 we show how different values of parameter β influence the time needed for user's reputation to fall on value $I_n = 1$ due to user's inactivity and value of dynamical reputation in the moment of the last activity. The higher the value of parameter β and initial dynamical reputation of users, the longer time it takes for user's reputation to fall on baseline value. For parameter $\beta = 0.9$ and $I_n = 5$, user's reputation falls on value $I_n = 1$ after less than 20 days, while this time is doubled for $\beta = 0.96$. We see, that for higher values of parameter β the time needed for I_n to fall on value 1 becomes longer, and that the the initial value of reputation becomes less important.

Figure A2 in SI shows the difference between the number of users that had at least one activity in the window of 30 days and number of users with reputation higher than 1 during the same period for different values of parameter β. The minimal difference between these two variables is observed for the values of β between 0.94 and 0.96 for both live and closed communities. Since we want to compare communities, we select $\beta = 0.96$ after verifying that this level of reputational decay does

not reduce the number of active users (based on their dynamic reputation) below the actual number of users who have been active (interacted with the community) in the time window of 30 days.

B.2 Dynamic reputation - β parameter

Our implementation of dynamic reputation model was based on $\beta = 0.96$. There are several reasons for selecting this value.

In Dynamic reputation model, the β parameter controls the strength of the forgetting fator of the model. The value of this parameter should reflect the core feature of the reputational systems and make reputation easier to loose. Due to user's inactivity, any level of reputation will eventually decay to below 1. Dependence of time needed for reputation to drop below this level and the β parameter, as well as reputation before inactivity is shown on Figure **??**. Here I_n is equal to the raw number of interactions in the community without forgetting or cumulative factor at work.

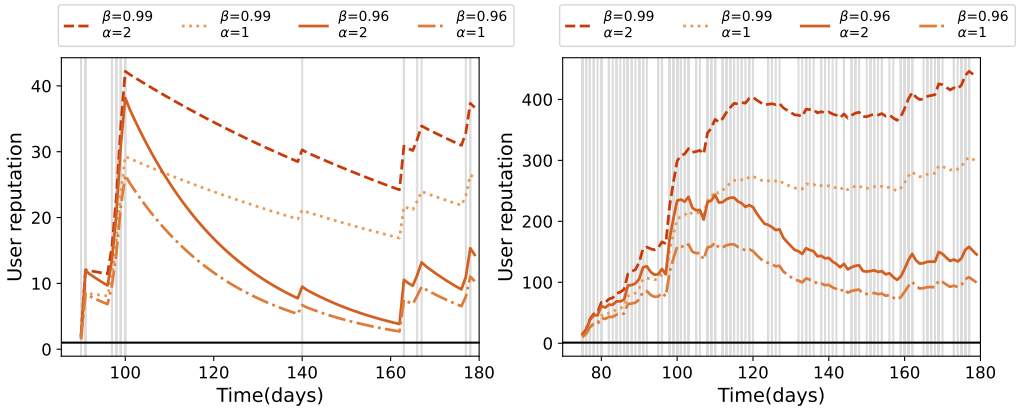

Figure B.1: Single users reputations

For β values below 0.96, the decay is fast and within two to four months of inactivity even high values of reputation are reduced below the threshold. On the other hand, with β values the decay proces is more differentiated and high reputation becomes harder to loose, surviving up to a year of inactivity. For β equal to 0.96, it takes a month for reputation based on 5 interactions to decay and around five months for high reputation based on 500 or 1000 interactions to decay below the threshold.

30 days sliding window We compared the number of users with estimated reputation higher than 1 for different parameters β and concluded that β close to 0.96 approximates the number of users with recorded interactions in a given 30 days sliding window. For each pair of communities we calculated number of users with at least one interactions in every 30 days sliding window and then we estimated several times series expressing the number of users with reputation higher than 1 for fixed β. Then we calculated the root mean square error (RMSE) between those time series for the first 200 days. Values of RMSE are shown on Figure B.2. For each community, we can find parameter β that minimizes RMSE. Although β does not have a unique value across communities, it varies between 0.95 and 0.96.

Figure B.3 shows comparison between number of users in 30 days sliding window, number of users for these optimal values $\beta = 0.954$ and $\beta = 0.96$. For $\beta = 0.96$ we observe that in most communities estimated number of active users consistently slightly higher than the actual number of users which have made at least one interaction in that sliding window. This means that dynamic reputation model in some cases overestimates the reputation of the user, but far more important is that it never

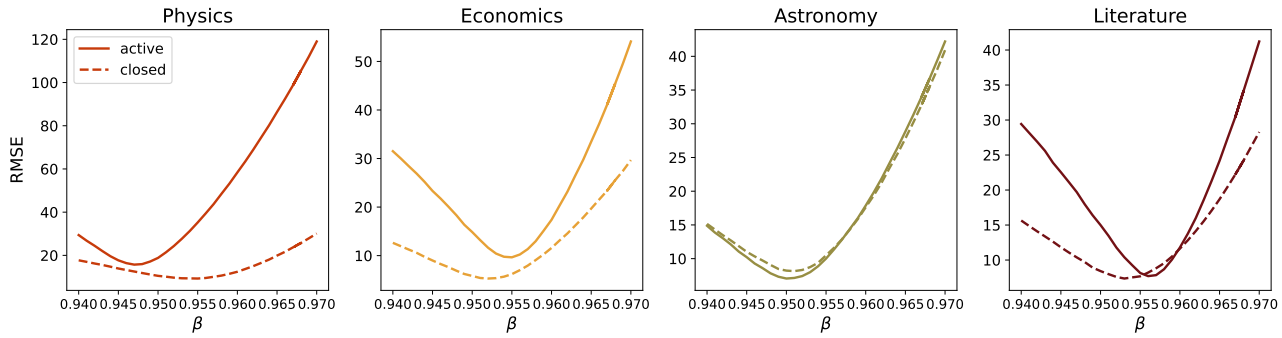

Figure B.2: RMSE between number of active users in sliding window of 30 days and number of users with reputation > 1 for $0.94 < \beta < 0.97$ with step 0.001.

understimates the real number of active users. Since we base our calculations of total and average reputation within the community only on users whose reputation is higher than the threshold this is important as no active users are disregarded by the model due to the value of the decay parameter.

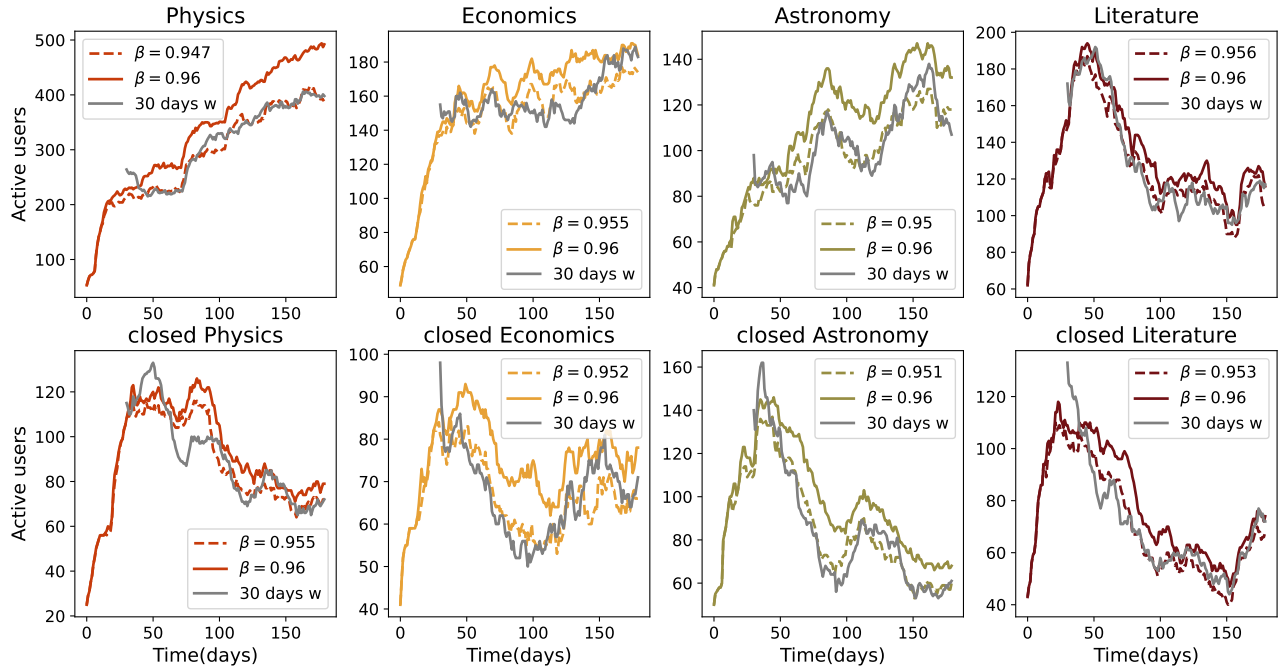

Figure B.3: Number of active users in a sliding window of 30 days and number of users with dynamic reputation higher than 1 for $\beta = 0.954$ and $\beta = 0.96$ which provide the best fit to the number of users in 30 days sub-networks for each community

Finally, it's imporant that our dynamic reptuation captures the trend of long-term user activity. In Figure B.4 solid lines show the time series of estimated dynamic reputation for $\beta = 0.96$ while dashed lines show the number of users who were active in a given sliding window and continued to be active in the next one. Although the total estimated number of active users is expectedly higher, two time series follow similar trends in different communities.

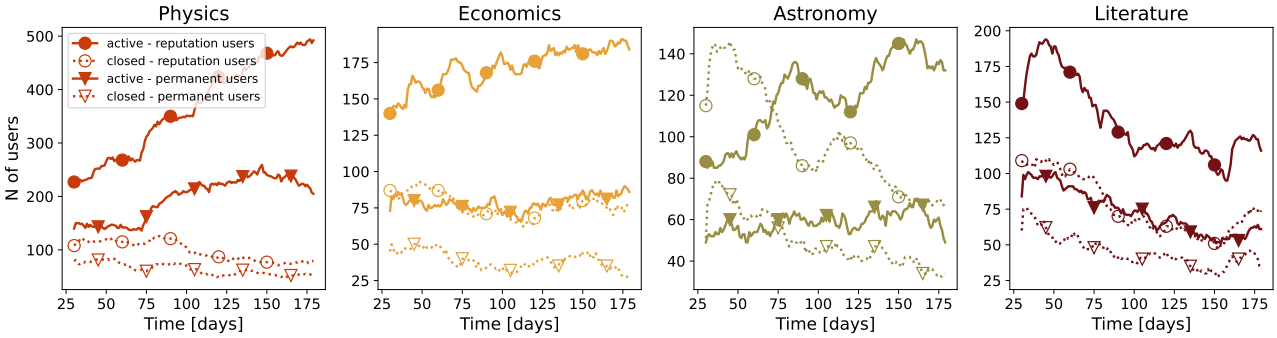

Figure B.4: Solid lines represent number of users with dynamic reputation higher than 1 for $\beta = 0.96$ while dashed lines are number of users within 30 days sliding window who were active and remained to be active. Blue lines are beta, while red lines are area51 communities.

Appendix C

The choice of the sliding window

There is no well-specified procedure for the choice of sliding window τ. Previous studies showed that if τ is small sub networks become sparse, while for too large sliding windows some important structural changes can not be observed [11, 12]. We analyse how networks properties and properties of dynamical reputation change with the window size, see SI for more details. Figure A13 in SI shows how considered network properties and dynamical reputation depend on the time window size for active and closed communities on the astronomy. We observe that fluctuations of all measures are more pronounced for time window of 10 days than for 30 and 60 days. However, we find that while the structural properties of networks evolve at different paces over varied time windows the trends remain very similar. The observed qualitative difference between closed and live communities is independent of τ, especially if we compare time window size of 30 and 60 days. The time window size of 30 days ensures enough amount of interaction, even for closed communities, while the number of observation points remains relatively high. For these reasons, we choose a sliding window of 30 days.

In this section, we investigate how the size of sliding windows affect network properties over time. Figure C.3 summarize results for one pair of communities, area51 and beta astronomy, but similar conclusions can be observed for other pairs of sites. We show the network properties for sub-networks of 10, 30, and 60 days sliding windows. For a sliding window of 10 days, results may be too noisy and we may not observe some important trends in the community. The number of users for beta astronomy seems to fluctuate around some mean value. On the larger scale, 30 days window, it is more clear that the number of users slightly increase over time. Contrary, for too large an aggregation window (60 days), important information about the time series can be lost, such as the local minimum of the number of users around time step 80 that is observed for the 30-day sliding window. Looking into other network characteristics such as L/N and clustering we conclude that differences between closed and active sites are more transparent with a larger aggregation window, still, on each scale, beta sites show a higher number of nodes, number of links per node and clustering coefficient.

As before we study the structure of created sub-networks through the lens of core-periphery structure. On small scales, the window of 10 days, there are often few, or even no nodes in the core and it can affect the calculation of other measures of interest. Such behaviour is more typical for closed communities. With the size of the sliding window number of nodes in the core increases and results of core-periphery measures become smoother. Finally, the choice of the sliding window does not change conclusions that core users in the beta communities produce more activity and make the strong core. However, our main results are shown for a sliding window of 30 days, as it makes a good compromise

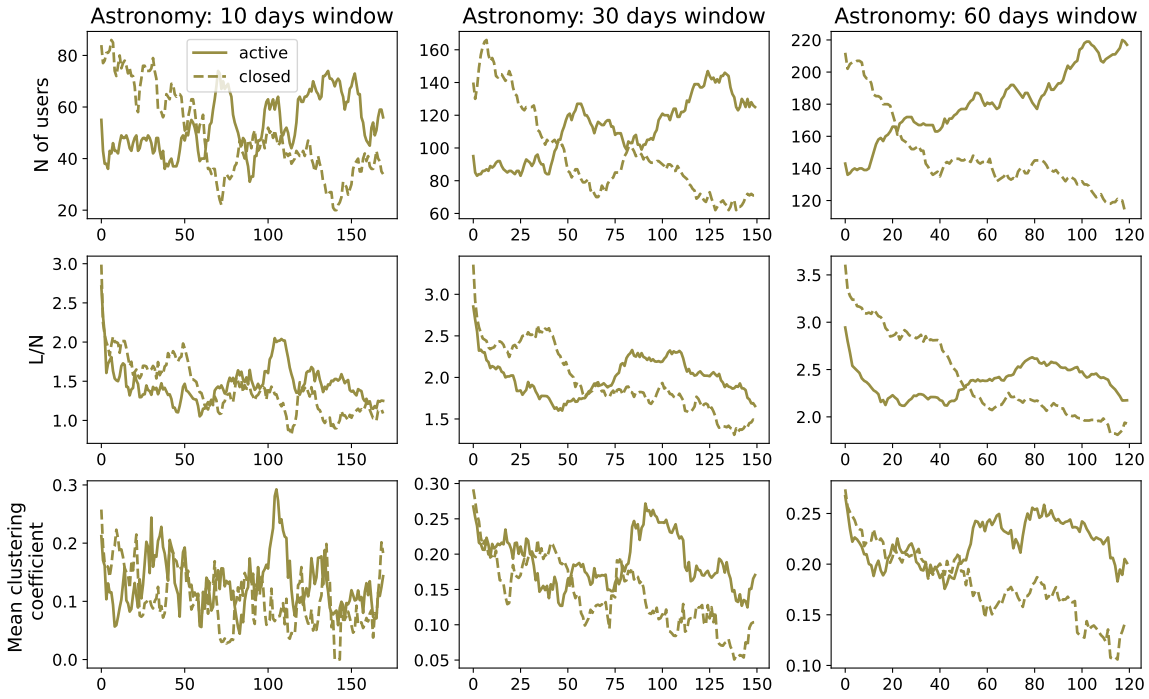

Figure C.1: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

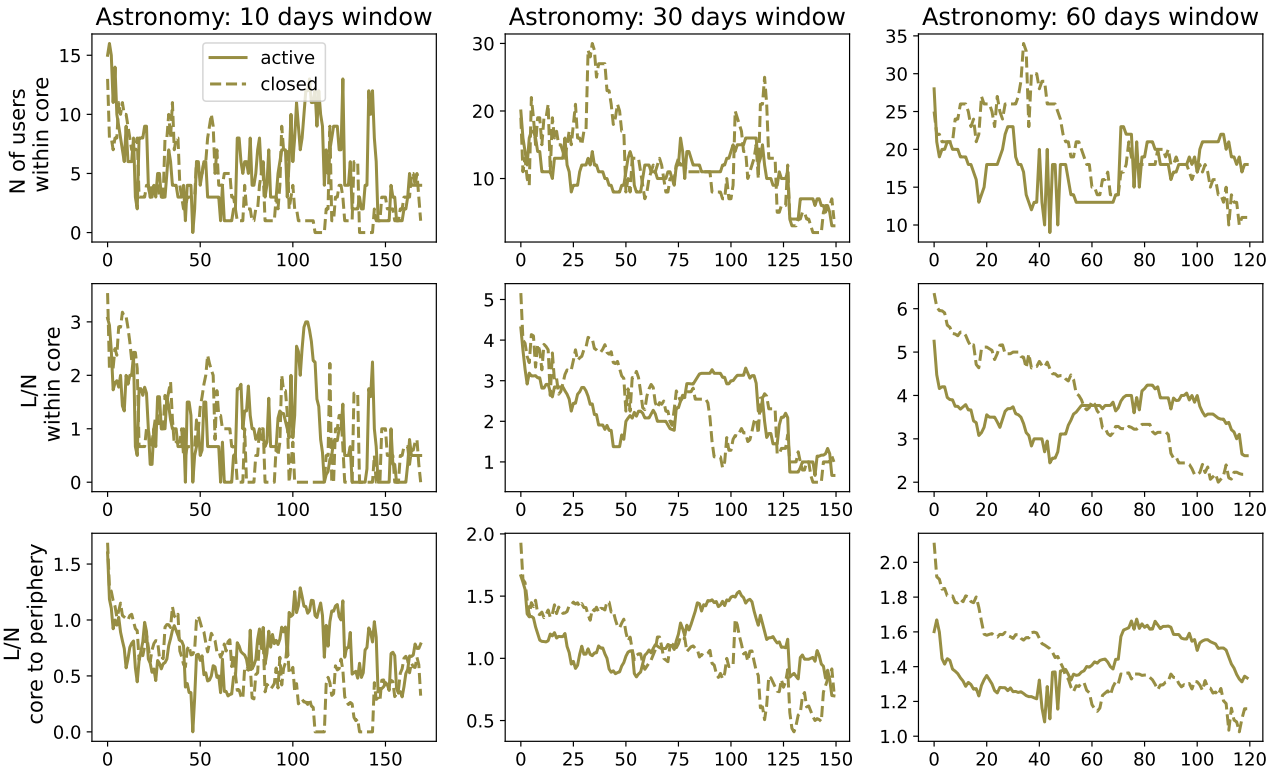

Figure C.2: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

between large and small time scales.

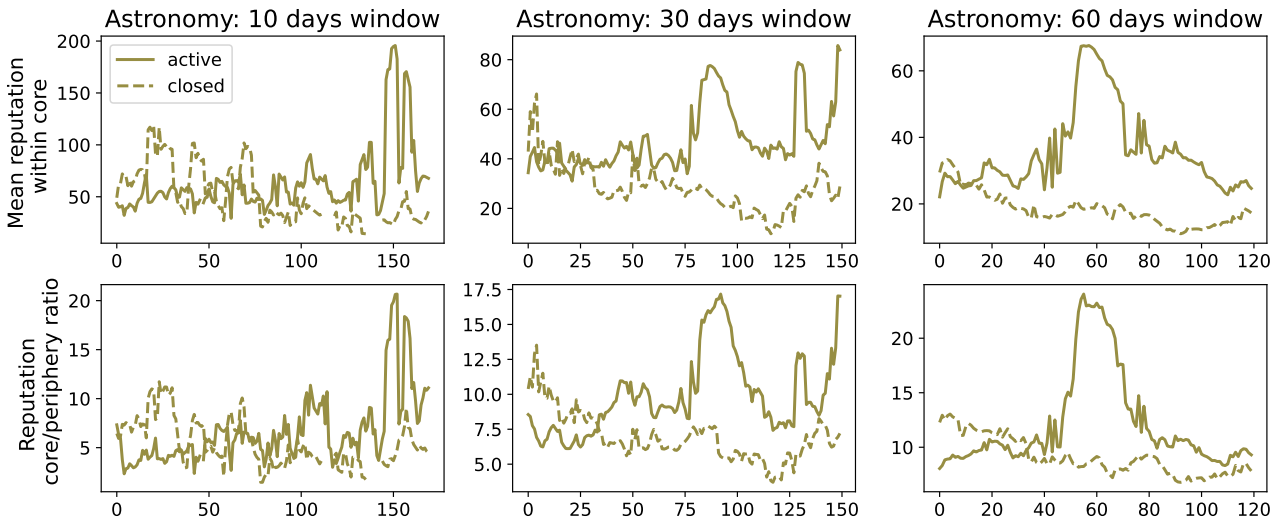

Figure C.3: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

Appendix D

Robustness of core-periphery algorithm

D.1 Precision and recall

Consider the network $G(V, L)$, with a set of nodes V and a set of links between them L. The stochastic community detection algorithms may converge to different configurations. To quantify the similarity between the obtained structures and robustness of the algorithm, we run 50 iterations and calculate several similarity measures between pairwise partitions C and C'.

The core-periphery structure has two groups so confusion matrix [55] can be defined as:

		partition C	
		core	periphery
partition C'	core	n_{TP}	n_{FN}
	periphery	n_{FP}	n_{TN}

The diagonal elements correspond to the number of nodes found in the same class in both node configurations. The number of nodes in the core found in C and C' is denoted as true positive n_{TP}, while the number of nodes in the periphery in C and C' is denoted as true negative n_{TN}. The off-diagonal elements of the confusion matrix indicate the number of nodes differently classified. We can define the number of nodes found in the first configuration C in the core but in C' in the periphery as a false positive, n_{FP}, similarly the number of nodes found in the periphery in the partition C, and in the core in partition C' as a false positive, n_{FP}.

From the confusion matrix, we can write the precision $P = n_{TP}/(n_{TP} + n_{FP})$ and recall $R = n_{TN}/(n_{TN} + n_{FN})$. These measures range from 0 to 1. The precision (recall) corresponds to the proportion of instances predicted to belong (not belong) to the considered class and which indeed do (do not) [55].

D.2 F1 measure

The **F1 measure** is the harmonic mean of precision and recall [55]:

$$F_1 = 2\frac{P \cdot R}{P + R} = \frac{2n_{TP}}{2n_{TP} + n_{FN} + n_{FP}} \tag{D.1}$$

It can be interpreted as a measure of overlap between true and estimated classes; it is 0 for no overlap to 1 if overlap is complete.

D.3 Jaccard coefficient

The **Jaccard's** coefficient is the ratio of two classes' intersection to their union [55]. It can also be expressed in terms of confusion matrix:

$$J = \frac{C_{core} \cap C'_{core}}{C_{core} \cup C'_{core}} = \frac{n_{TP}}{n_{TP} + n_{FP} + n_{FN}} \tag{D.2}$$

D.4 Normalized mutual information

Normalized mutual information (NMI) is similarity measure between to partitions C and C' based on information theory [56]:

$$NMI(C, C') = \frac{MI(C, C')}{(H(C) + H(C')/2} \tag{D.3}$$

where MI is mutual information between sets C and C', while $H(C)$ is entropy of given partition. The entropy is defined as $H(C) = -\sum_{i=1}^{|C|} P(i)log(P(i))$, where $P(i) = |U_i|/N$ is the probability that an object is randomly classified as i (in this special case $i = 0$, the node belongs to the core, or $i = 1$, the node belongs to the periphery). The mutual information between sets C and C' measures the probability that the randomly chosen node is a member of the same group in both partitions:

$$MI(C, C) = \sum_i \sum_j P(i, j)log(\frac{P(i, j)}{P(i)P'(j)}) \tag{D.4}$$

where $P(i, j) = |U_i \cap U_j|/N$

NMI ranges from 0 when the partitions are independent to 1 if they are identical.

D.5 Adjusted rand index

Adjusted rand index. For the set of nodes V, with n nodes, consider all possible combination of pairs (v_i, v_j). We can select the number of the pairs where nodes belong to the same group in both partitions, C and C', denoted as a. Similarly, as b, we can define the number of pairs whose nodes belong to different groups in partitions. Then, unadjusted rand index [57] is given as $RI = \frac{a+b}{\binom{n}{2}}$,

where $\binom{n}{2}$ is number of all possible pairs. The RI between two randomly assigned partitions is not close to zero; for that reason, it is common to use the adjusted rand index [58] , defined as:

$$ARI = \frac{RI - E[RI]}{max(RI) - E[RI]} \tag{D.5}$$

where $E[RI]$ is expected value of RI, and $max(RI)$ is maximum value of RI.

As example we show analysis of inferred sample of core-periphery structures for 30 days closed Astronomy, Stack Exchange networks, Figure D.1. We represent the mean minimum description length (MDL) and the mean number of nodes in the core with standard deviation. MDL does not change much between inferred core-periphery structures; the difference between obtained configurations is still notable in the number of nodes in the core. To investigate in more details similarity between obtained core-periphery configurations in the sample we calculate several measures between pair-wise partitions such as normalized mutual information, adjusted rand index, F1 measure and Jaccard index. These measures are greater than 0.5 and, in most cases, greater than 0.9, indicating stability of the inferred core-periphery structures.

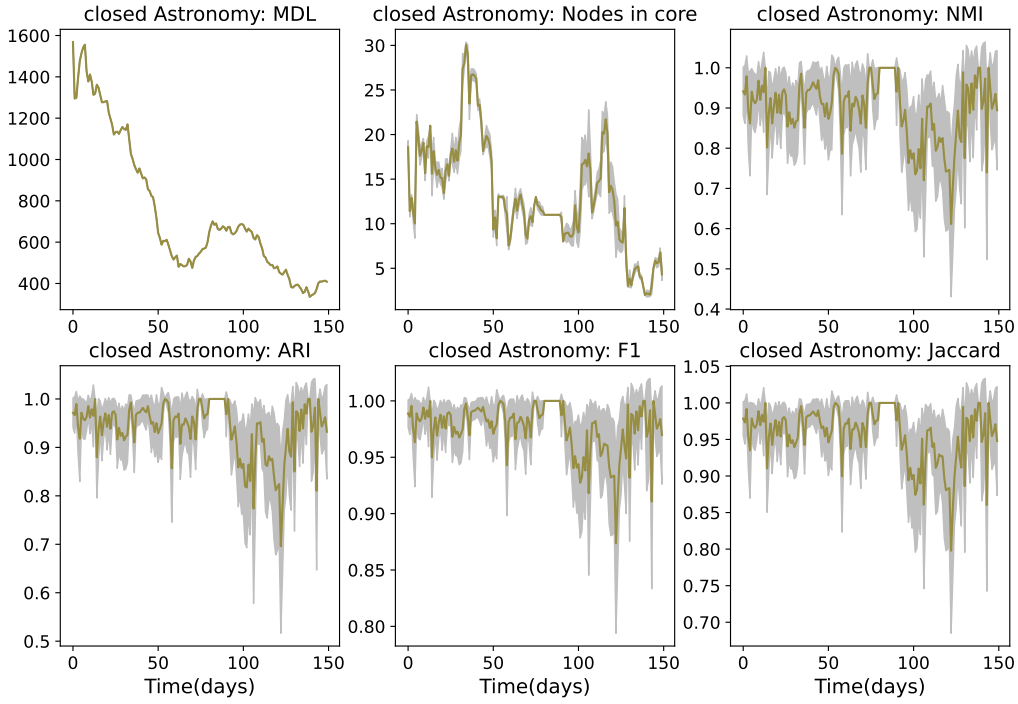

Figure D.1: Minimum description length, number of nodes in core, normalized mutual information, adjusted rand index, F1 measure and Jaccard index, among 50 samples for 30-days sub-networks. Results are given for closed astronomy.

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] V. Latora, V. Nicosia, and G. Russo. Complex networks: Principles, methods and applications. 2017.

[3] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007.

[4] Luciano da Fontoura Costa, Osvaldo N Oliveira Jr, Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.

[5] Guido Caldarelli. *Scale-Free Networks: Complex Webs in Nature and Technology*. 05 2007.

[6] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016.

[7] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[8] Petter Holme and Jari äki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[9] Naoki Masuda and Renaud Lambiotte. *A Guide to Temporal Networks*. 10 2016.

[10] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):1–30, 2015.

[11] 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.

[12] 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.

[13] Mason A Porter. What is... a multilayer network. *Notices of the AMS*, 65(11), 2018.

[14] Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10(1):45–62, 2019.

[15] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 07 2014.

[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] 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.

[18] Martin Rosvall, Jean-Charles Delvenne, Michael T. Schaub, and Renaud Lambiotte. Different approaches to community detection. *CoRR*, abs/1712.06468, 2017.

[19] 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.

[20] Kamalika Basu Hajra and Parongama Sen. Phase transitions in an aging network. *Physical Review E*, 70(5):056103, 2004.

[21] Thorben Funke and Till Becker. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE*, 14(4):1–40, 04 2019.

[22] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[23] J. Nair, A. Wierman, and B. Zwart. *The Fundamentals of Heavy Tails*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2022.

[24] 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.

[25] Espen Alexander Fürst EAFI Ihlen. Introduction to multifractal detrended fluctuation analysis in matlab. *Frontiers in physiology*, 3:141, 2012.

[26] 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.

[27] Sergey N Dorogovtsev and José FF Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63(5):056125, 2001.

[28] Jelena Smiljanić and Marija Mitrović Dankulov. Associative nature of event participation dynamics: A network theory approach. *PloS one*, 12(2):e0171565, 2017.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] Gerald F Frasco, Jie Sun, Hernán D Rozenfeld, and Daniel Ben-Avraham. Spatially distributed social complex networks. *Physical Review X*, 4(1):011008, 2014.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] Jop Briët and Peter Harremoës. Properties of classical and quantum jensen-shannon divergence. *Phys. Rev. A*, 79:052311, May 2009.

[42] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[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] Akrati 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.

[45] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Activity archetypes in question-and-answer (q8a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing*, 2(1):1–23, 2019.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.

[53] 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.

[54] 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.

[55] Vincent Labatut and Hocine Cherifi. Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*, 2012.

[56] 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.

[57] 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.

[58] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.