UNIVERSITY OF BELGRADE
FACULTY OF PHYSICS

Ana Vranić

EVOLVING COMPLEX NETWORKS: STRUCTURE AND DYNAMICS

Doctoral Dissertation

Belgrade, 2022

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЗИЧКИ ФАКУЛТЕТ

Ана Вранић

РАСТУЋЕ КОМПЛЕКСНЕ МРЕЖЕ: СТРУКТУРА И ДИНАМИКА

докторска дисертација

Београд, 2023. година

Thesis Defense Committee

Thesis advisor:

Committee members:

Acknowledgements

Abstract

Complexity science gives us new ways to explore complex systems. Detecting the collective phenomena and understanding how they emerge from individual interactions is one of the important research problems. Complex systems are all around us; they come from physical, biological, and social systems. The complex network is a general framework for representing interaction patterns in complex systems. The structure of the network could influence the behaviour of the system. The discovery that real-world networks are far from random and show scale-free properties and small-world phenomena led to the development of this field. New theories and models were needed to study socio-economics phenomena. At the same time, statistical and computational physics tools helped analyse and model their complex network representations. This thesis aims to broaden the knowledge of complex networks by empirically analysing different online social systems and providing the models and theories that could explain their specific characteristics.

The first part of the research explores how different growing signals influence the structure of complex networks. Over time, systems do not grow at a constant rate, and the networks that grow under fluctuating signals are clustered and correlated, while networks grown with a constant signal are not. Here, we systematically understand the connection between the growth signal and the network structure. For the growing network model, we use time series of new users from natural systems, such as MySpace and TECH. At the same time, computer-generated long-range correlated signals help distinguish which properties of time series shape the structure of complex networks. When signals are correlated and have multifractal properties, they mainly influence the scale-free networks promoting the creation of highly connected nodes.

The second part of the research focuses on the evolution of large online platforms, where users organise into different kinds of social groups. These days, people interact intensively through online platforms. No matter whether the online systems rapidly grow, universal patterns in their growth stay stable. Our approach was to empirically analyse the evolution of three online systems: Meetup groups in London and New York and Subreddits. Their group size distributions follow log-normal, indicating the presence of universality. On the other hand, it was important to identify the processes that led to the emergence of log-normal distribution and provide a model that could produce growth patterns in real systems. Social connections could be an important factor in the diffusion between groups. We used a model that interplays two criteria for group choices: random and based on social connections. We showed that social interactions are more critical in Subreddits than in Meetups for the diffusion between groups.

The last part of the research, presented in the thesis, addresses what is necessary for one community to be sustainable. The complex network representation of the system allows us to determine how different network properties evolve. We use data from Stack Exchange sites, comparing communities on the same topic, but one was closed, and later when the site was proposed later, it stayed active until

these days. Stack Exchange sites are question and answers platforms where users share knowledge. Analysing the structural patterns in these communities, we found active ones to be more clustered and characterised by better-connected cores. Core users are crucial for a healthy site and need to be trustworthy. Through the dynamic reputation model, we attempt to measure the level of trust in these communities. In active communities, core users show a higher reputation than in closed communities, indicating the importance that a stable core develops early and has a high level of trust. **Keywords:**
Research field: Physics
Research subfield: Statistical physics
UDC number: 536

Сажетак

Кључне речи: Научна област: Физика
Ужа научна област: Статистичка физика
УДК број: 536

Contents

List of figures

List of Tables

Chapter 1

Introduction

Many real systems, such as brain networks, social organisations, cities or even biological systems, can be represented as complex systems. The common property of these systems is that they are composed of many interacting elements. Still, to describe the properties of the system, we can only conclude a little from the behaviour of a single individual. Due to specific interactions, without any central force, in the complex system, the collective behaviour [1] emerges. The structure of the brain network and its properties are fundamental for brain functioning, while an emergent phenomenon is human intelligence. In societies, people's interactions lead to civilisation, economy, formation of social groups. Also, the animal populations show different levels of organisation: such as patterns in bird flocks or schools of fish [2].

The interactions between the complex system elements are not homogeneous; as systems evolve, they can also change [2]. The research in complex systems mainly focuses on the interactions between its units. Knowing the shape of these connections, we can determine the properties of the system [13]. We can construct a representation with neurons and synapses representing connectivity in the brain network. Neurons in the same brain area are closely connected [14]. 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.

The universality is an important property of the complex systems [3]. For example, the time lap between two email messages follows the power-law distribution [4], and the exponent is universal across different platforms. Similar conclusions are found in distributions of the votes in elections [5, 6], and citations of scientific publications [7]. Even the growth of social groups, such as cities, follows universal patterns. The probability distribution of the city sizes in one country follows the same laws, with a similar exponent for all countries [8, 9]. However, the distribution of company sizes follows log-normal behaviour and remains stable over decades [10, 11]. Understanding how universality emerges in different systems is the focus of the statistical physics of the complex systems [12].

Despite the differences between complex systems, they can be studied using the same techniques. The natural extension of the complex system is the network, 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 [15, 16] (graph of actors), biological (network of proteins) [17, 18] or even technological systems (internet, traffic) [19, 20, 21].

The complex network theory originates from the graph theory in mathematics. The first problem solved using graph theory was the $Konigsberg$ problem of seven bridges. The city of $Konigsberg$ had seven bridges connecting the city's parts across the river and the island in the middle. Is it possible to find a walk that crosses all seven bridges only once? Representing the problem as a graph, Euler managed to simplify the problem; the parts of the land are represented as nodes while bridges between them are links, see Figure 1.1. Crossing each bridge only once is possible if each part of the land has an even number of connections. It makes it possible to enter one part of the land from one bridge and leave it on the other. As each node has an odd number of connections, it is impossible; see Figure. 1.1.

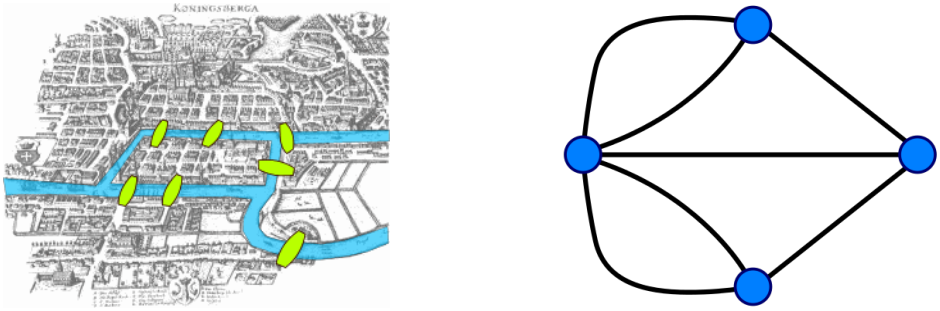

Figure 1.1: The Kronigsber problem of seven bridges. The left panel shows the original map of the bridges; the right panel shows its graph representation.

Until the late 1990s, graph theory was not widely used. Back then, the most crucial model was the Erdos-Reni model of random graphs, which considers a fixed number of nodes in the network connected randomly, resulting in the Poisson degree distribution. When researchers got an idea to map the World Wide Web (WWW) on the network and analyse its properties [22], they found that degree distribution follows the power-law contrary to expected behaviour from random graph model [23]. Because the power-law distribution is the same on all scales, such networks are called scale-free. Besides the scale-free property, empirical analysis of various complex networks showed the small-world property, and the high clustering coefficient [24, 25]. Two seminal papers from 1999 inspired further research in complex networks. Watts and Strogatz [26] proposed the model where rewiring of edges on regular lattice leads to the network in which paths between any two nodes become small (small-world) and nodes become densely connected, resulting in a high clustering coefficient. On the other hand, Barabasi and Albert (BA) [27] introduced the model, where the network grows over time, and the new nodes tend to connect high-degree nodes; it produces scale-free networks with few highly connected nodes.

Different complex network models were proposed to describe the structure and dynamics of social and technological systems. The node degree is one of many node features that determine the linking probability, and the linking probability may be nonlinear in node degree or may depend on the age of the node [28, 29]. In the BA model, the links are introduced through new nodes, so it was proposed that links can be created between existing nodes in the network.

Furthermore, the BA model considers the constant network growth, where a fixed number of nodes is added at each step. The research on various social systems shows time-dependent growth, and we record the exponential growth of online systems [30]. Some models considered that nodes become inactive or even that network grows through a nonlinear number of links [31]. On the other hand, models with accelerated growth in the number of nodes [32] simulate exponential expansion of the online social systems. But the growth is not only accelerated; the time series of new nodes has trends and reflect the typical human behaviour [33, 34, 35].

Research has also been devoted to using generated networks to analyse dynamic processes on top of them. Central questions are about the spread of epidemics, information diffusion, or emotional inter-

actions among elements [4]. These systems are modelled using agent-based models, while the robustness is often studied by percolation and diffusion phenomena in complex networks. It was shown that scale-free networks' connectivity is sensitive to removing highly connected nodes. On the other hand, eliminating small degree nodes won't affect the scale-free structure [36]. They also show resilience to random attacks. Real-world networks are often characterised by community structure. They are common for social networks, where people with similar interests group together. Mostly adopted definition of a community is a group of densely connected nodes. The complex network theory provides different models for generating networks with community structure but also develops the algorithms for inferring the community structure from the underlying network.

The complex network models contribute to our knowledge, connecting the network topology and the dynamics of the system and helping us to understand underlying mechanisms that lead to the emergence of the properties of the complex networks [27, 37, 38, 39]. Complex network models must gain insights based on empirical data and social theories, and they are data-driven and require the development of computational approaches. The physicists showed interest in modelling complex systems by applying statistical physics approaches. Recently, the theory of graph neural networks (GNN) emerged from computer science, where machine learning methods are found helpful in inferring the properties of the network. For example, they are used to determine missing links and recommend to users in online social networks or to develop generative GNN models that lead to the discovery of new drugs.

Real networks are much more heterogeneous than networks obtained in simple models. Links may be directed or undirected, they may have temporal dependencies, or we can deal with different types of interaction in one system. Other network representations deal with these specific features. In the following section, we will introduce complex networks and different approaches to deal with particular data types.

1.1 The complex 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 the network is un-directed, edges are symmetric, so (v_i, v_j) implies (v_j, v_i). In **directed** networks, this symmetry 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 a directed network. The first column a) in Figure 1.2 shows the graphical representation of two networks with an equal number of nodes; the first is undirected, and the second is directed.

Even though graphical representation can be useful for describing the network structure, mathematical representation allows us to characterise the statistical properties of the networks. The graph G, with N nodes could be represented with **adjacency matrix** $|A| = N \times N$ [40]. The matrix elements are positive if there is a connection between two nodes v_i and v_j.

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

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

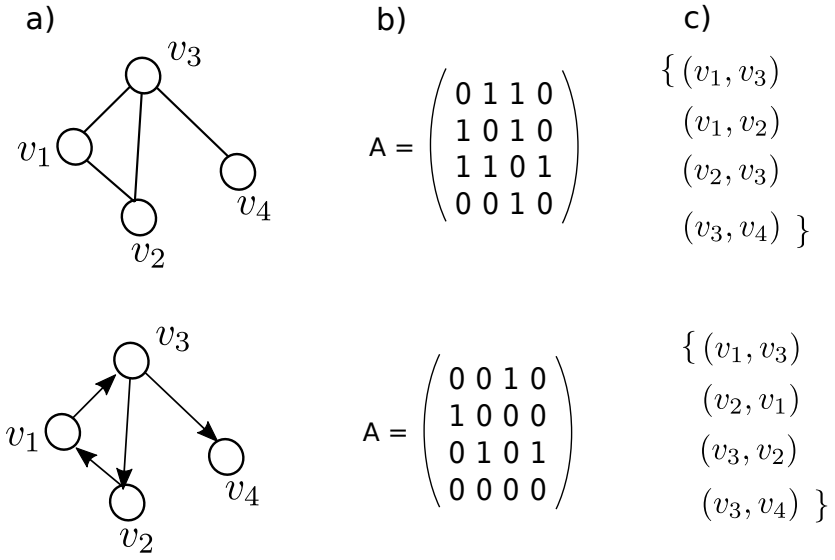

Figure 1.2: a) Graph representation of undirected (top panel) and directed (bottom panel) network. The same networks are represented with adjacency matrices in column b) and edge list representation in column c).

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 a directed network, it is possible to draw $L_{max} = N(N-1)$ edges [41]. When it comes to large networks, they are sparse, meaning that the number of links is $L << L_{max}$. Consequently, the adjacency matrix is also a sparse structure (has many zeros) that takes a large portion of computer memory [42]. It is common to represent the graph as an edge list. In this case, illustrated in Figure 1.2, column c), a graph is described with the list of links that are in the graph, $G = \{\{v_i, v_j\}\}$. Still, with this representation, we cannot distinguish between directed and undirected graph structures, so the computational algorithm should specify if the edges are symmetric or not.

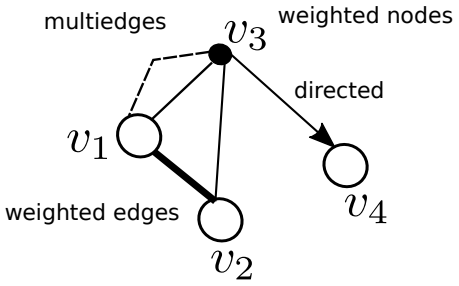

Figure 1.3: The complex networks may represent different system characteristics. The edges can be directed, weighted or multiply. Also, nodes can be assigned with different weights or any relevant feature.

Sometimes is essential to include the specific properties of the system in the network representation. For example, to emphasise the frequent interactions between nodes, edges can be assigned with different values; such networks are **weighted**. They can be described with an 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, properties can be added to nodes or the whole network structure. Edges could be characterised by the time when the interaction between nodes happens, which includes the **temporal** component in the network representation. Finally, if two nodes interact differently, the **multigraph** is an appropriate configuration where multiple edges are allowed. Figure 1.3 presents the graphical representation of discussed network representations.

A **bipartite network** consists of two types of 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[42, 14]. 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 user and item link is created when the user buys an item. The bipartite networks find their application in the algorithms for recommended systems, whose goal is to suggest 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 [25, 42].

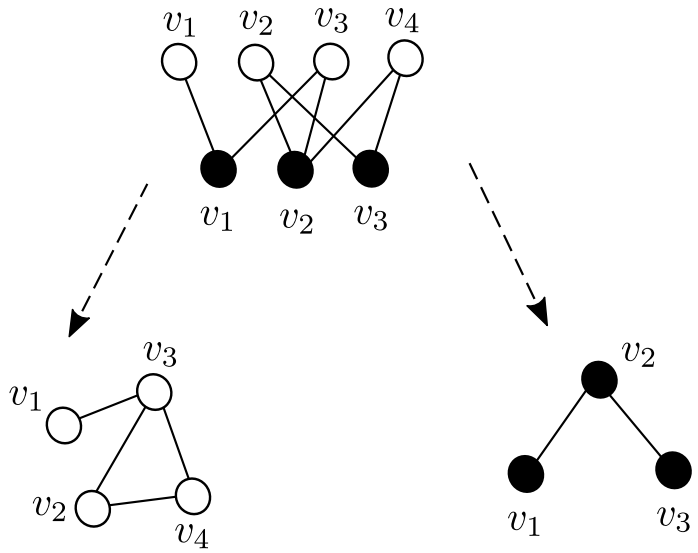

Figure 1.4: Bipartite network and two partition projections.

We should be aware that important information is lost when creating a one-mode projection. First, having weighted edges in the network of actors is necessary to know how many movies two actors appear. 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, i.e. 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 real systems as static networks can give us a lot of insight into the system's 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 [43]. For example, if the 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 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 [44]. The duration of the events may vary, as in the phone-call network. Also, for many systems, the time resolution of the 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.

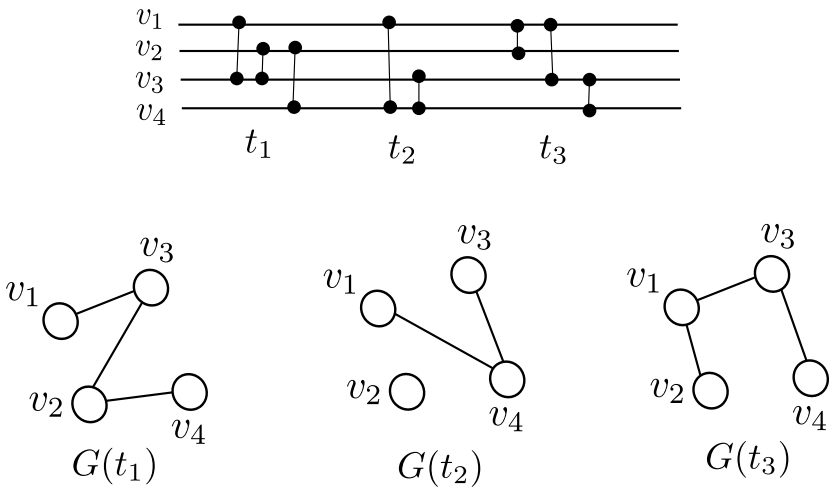

Figure 1.5: Temporal network.

The temporal network can be represented as a 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 analyse 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 [45]. 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 following 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 [46, 47].

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 [48]. 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 [49]. Or the airline transportation network, where each layer represents the flights of different airline companies [50].

1.2 In this thesis

This thesis uses statistical physics and complex network approaches to model and empirically analyse online social systems. These systems consist of many users interacting online and could be represented by complex networks. In chapter 2, we provide the methodology employed for this research. We describe the fundamental measures of complex networks and introduce basic complex network models. We review the most common probability distributions characterising complex systems' properties and outline distribution fitting methods. Finally, we introduce the multifractality of the time series and dynamical reputation model.

Chapter 3 addresses the difference between network models where the growth in a number of nodes is constant and when it follows a non-trivial growth signal. This research aims to quantify how growth signals influence the structure of complex networks. Using the adapted ageing model [51], we use computer simulations to generate different kinds of complex networks. For more realistic real-world network simulations, growing signals are time series of new users from online social platforms, MySpace and Tech group from Meetup. They are described with trends, cycles and long-range correlations. Often time series have multi-fractal properties. The results of this study are published in [52], and they show the importance of growth signals in shaping the network structure because the scale-free networks, which represent real systems, are mainly altered.

As research on social groups mainly focuses on a single group, there are remaining questions about the characteristics of the entire system. For example, the Tech group is only one of the groups around which Meetup users organise; many other groups are created worldwide, so the system constantly grows. In chapter 4, we will examine how groups on online social platforms grow. The results are summarised in the paper [53]. This research is based on Reddit and Meetup data. From Meetup, we created two data sets, one with groups created in London and the other with groups created in New York, while for Reddit, we selected groups built before 2012. We are interested in explaining scaling behaviour in group size and growth rate distributions and identifying the growth mechanisms present in the system. Using a bipartite complex network model, we can reproduce the universality found in the system.

Even though across complex systems, we find the emergence of universal behaviour, for example, the scaling of the degree distribution of two groups is similar, different factors might influence its success. It is well known that many online groups may suddenly fall apart. These questions are the subject of the chapter 5, which main results are published in the paper [54]. Here, we study the question-answer platform Stack Exchange; it has more than 200 different topic-specified sites where people help each other answer questions. What is interesting about this system is that some sites were closed because they did not produce enough activity. For that reason, we selected the sites with the same topic that failed but later, when someone proposed the site again, it stayed active. We analyse the evolution of user interaction networks; here, we use the temporal network approach and compare active and closed sites. We find that it is essential how the network users are distributed into a core-periphery structure [55]. The core must select firmly connected users, but their interaction with the periphery has to be high. In other words, we need a trustworthy core to hold the community. Introducing the Dynamical Reputation Model (DIBRM) [56], based on user interaction sequences, we quantify how much users can be trusted and whether the community has a strong core. In the appendix A, we briefly describe the Stack Exchange sites. In appendix B and C discuss how we choose parameters for the DIBRM model, while in appendix D we discuss the stability of inferred core-periphery structures.

Finally, in chapter 6, we draw the main findings of this thesis.

Chapter 2

Methodology

2.1 The measures of complex networks

The complex system can be represented by a complex network $G = (V, E)$, where the elements of a system (atoms, proteins, people) map to a set of N nodes $V = \{1, 2, ..., N\}$. The interactions between elements map to L links between nodes, $E = \{e_1, e_2...e_L\}$. The **adjacency matrix** $A = N \times N$ has value 1 if there is a connection between two nodes; otherwise, it is 0 [40]. There are a lot of measures to quantify the structure of the network. This section lists the most important measures, such as degree distribution, correlations, and shortest path measures. We also discuss different structures found in the network, such as core-periphery or community structures.

2.1.1 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 network density is the average degree divided by $N - 1$, where N is the number of nodes. It is a 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 a more complex structure. If the degree sequence is skewed, we can identify nodes with high-degree (hubs). Removing hubs may partition a connected network into several components.

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 a randomly chosen node has degree k. Similarly, we can order nodes according to their degree and plot the node degree.

If the graph nodes are statistically independent, the degree distribution completely determines the properties of a network [57].Here we summarize the forms of degree distributions that are mostly found in the complex network theory:

- The Poisson distribution. The degree distribution in a 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. $P(k) = e^{-k/k}$. It is the degree distribution of the growing random graph. Even for infinite networks, all moments of distributions are finite and have a 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. No natural scale exists in this distribution, so they are called scale-free networks. In infinite networks, all higher moments diverge. If the average degree of scale-free networks is finite, then the γ exponent should be $\gamma > 2$. Therefore, real networks have a scale-free structure with the emergence of the hubs [25].

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

2.1.2 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 [58]:

$$r(x,y) = \frac{\frac{1}{n}\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.1}$$

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

Using the correlation coefficient definition, we can define correlations 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 $\boldsymbol{d} = [d_1, ..., d_n]$

$$r_{deg}(G) = \frac{\sum_{i=1}^{n}\sum_{i=1+1}^{n}((d_i - \bar{d})(d_i - \bar{d})\boldsymbol{A}[i,j])}{\sum_{i=1}^{n}(d_i - \bar{d})^2} \tag{2.2}$$

An adjacency matrix allows us to calculate the correlations between neighbouring nodes. If two nodes are not connected $A[i,j] = 0$, the degree of correlation between them does not contribute 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 for connections to existing between similar degree nodes. The negative correlations indicate that large-degree nodes prefer to connect nodes with a small degree, disassortative networks. The average first neighbor degree k_{nn} can be calculated as $k_{nn} = \sum_{k'} k' P(k'|k)$. The P is the conditional probability that an edge of degree k points to a node with degree k. The norm is $\sum_{k'} P(k'|k) = 1$, and detailed balance conditions [40], $kP(k'|k)P(k) = k'P(k|k')P(k')$ [40]. If the node degrees are uncorrelated, k_{nn} does not depend on the degree; otherwise, increasing/decreasing function indicates positive/negative correlations in the network [59].

The Newman defined the assortativity [60] index r in slightly different way:

$$r = \sum_{kl} kl(e_{kl} - q_l q_k)/\sigma_q^2, \tag{2.3}$$

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

2.1.3 Clustering coefficient

The **clustering coefficient** is a measure describing the neighbourhood's structure. In networks, exist a tendency to form triangles or clusters [42]. It 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,

$$c_i = 2e_i/(k_i(k_i - 1)) \tag{2.4}$$

We can calculate the mean clustering coefficient by averaging it overall network nodes. It ranges from $\langle c \rangle = 0$ where connections between neighbouring nodes do not exist; the network has a tree structure. 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 [61]. A triangle at node v is a complete subgraph with three 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 the ratio of the 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.

2.1.4 Network paths

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

The **path** between two nodes [58], 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. We can easily calculate the number of walks between two nodes using the adjacency matrix. The A^2 gives us walks of length 2, the A^3, the number of walks of length 3, and so on.

The network is connected if it can define the path between every two nodes. 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 [62].

We can find different paths between two nodes in the network, but the most important one is the **shortest path** [58, 62]. 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 a 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 value. In contrast, the **diameter** defines the largest distance between nodes in the network [58]. 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 [58].

$$\bar{d}(u) = \frac{1}{|V|-1} \sum_{v \in V, v \notin u} d(u,v) \tag{2.5}$$

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

$$\bar{d}(G) = \frac{1}{|V|} \sum_{u \in V} \bar{d}(u) \tag{2.6}$$

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

2.1.5 D-measure

For each node i, we can define the distribution of the shortest paths between node i and all other nodes in the network, $P_i = \{p_i(j)\}$, where $p_i(j)$ is percent of nodes at a distance j from node i. The connectivity patterns can efficiently describe the difference between the two networks. To specify how much G and $G^{'}$ are similar we use D-measure [63]

$$D(G,G^{'}) = \omega \left| \sqrt{\frac{J(P_1,..P_N)}{log(d)}} - \sqrt{\frac{J(P_1^{'},..P_N^{'})}{log(d^{'})}} \right| + (1-\omega)\sqrt{\frac{J(\mu_G,\mu_{G^{'}})}{log2}} \tag{2.7}$$

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

$$J(P_1,..,P_N)) = \sum_{i,j} p_i(j) log(\frac{p_i(j)}{\mu_j}) \tag{2.8}$$

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

The first term in equation 2.7 compares local differences between two networks, and Jensen-Shannon divergence between N shortest path distributions $J(P_1,...,P_N)$ is normed with network diameter $d(G)$. The second part determines global differences, computing $J(\mu_G,\mu_{G^{'}})$ between mean shortest path distributions. The D-measure ranges from 0 to 1. The lower D-measure is, the more similar networks are, and structures are isomorphic for D-measure $D = 0$.

2.2 Community structure

Nodes can be organized into groups called communities. In social networks, communities indicate that people share some common interests, or in biological networks, we can find that genes or neurons with similar functions are grouped. Identifying these hidden blocks can lead to interesting insights into the network. However, the community detection problem does not give a precise characterization of what a community is. A standard definition of a community is densely connected subgraph [64, 65], meaning that nodes in one community tend to associate, creating the assortative connectivity pattern. On the contrary, nodes could be organized in disassortative communities, where connections between groups are denser.

The network with k communities could be represented using $k \times k$ matrix p. The diagonal elements of p indicate the density inside communities, while off-diagonal elements show the density between groups. Figure 2.1 [64] shows the matrix and networks for two communities. In the first example, (2.1 a), the diagonal elements have a higher probability, as in the classic definition of assortative community structure. In disassortative structure (2.1 b), more connections exist between two partitions than inside them, i.e. off-diagonal elements have higher probabilities. Bipartite networks can be represented as a disassortative network with two groups. The links exist only between communities. Figure (2.1 c) shows the core-periphery network. This network structure is composed of a core where nodes are well connected with itself and with the periphery. The connectivity inside the periphery is sparse. Finally, if there is no difference between connectivity inside and between groups, the concept of communities is lost. We can treat the whole network as a single community, where each node has the same connectivity probability, i.e. as Erdos Renyi random graph.

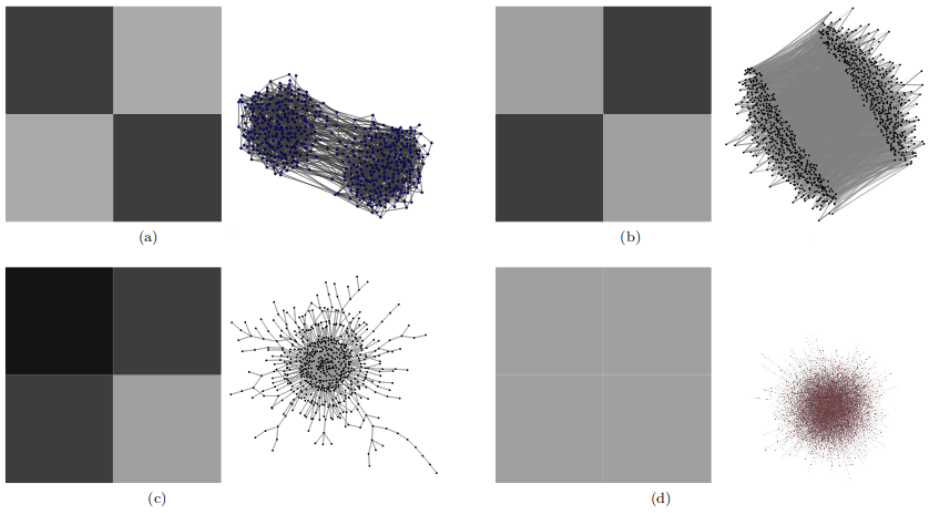

Figure 2.1: Different communities structures (a) assortative. (b) disassortative. (c) core-periphery. (d) Erdos Renyi random graph.

Different algorithms are used for detecting the community structure in the underlying network, optimizing different objective functions of the network partition. Still, if the ground-truth communities are unknown, there are no guarantees that we will infer the actual number of communities and entirely correct node assignments. Even though community detection algorithms are widely used in complex network analysis as they can give us a better understanding of network structure. In this section are explained two community detection models, the first one based on optimizing the modularity function, and the other based on the statistical inference of the Stochastic Block Model (SBM) where is optimized the likelihood function.

2.2.1 Louvain algorithm

Comparing the link density of the community with the link density obtained for the same group of nodes randomly connected, we could conclude if the community corresponds to the dense subgraph or if the structure is created entirely random. The **modularity** [66, 67] is a function that measures the randomness of each partition. We can compare the communities with modularity 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 the number of connections is larger than the expected number of links between N_c nodes given in the expected node sequence, 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 $p_{ij} = \frac{k_i k_j}{2L}$.

$$M_c = \frac{1}{2L} \sum (A_{ij} - p_{ij}) \tag{2.9}$$

If modularity is positive, the selected nodes may be a community, as their connectivity is far from random. If M_c is zero, then the connectivity between nodes is arbitrary, 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 the network partitioned into n_c communities is then defined as:

$$M = \sum_{c=1}^{n} [\frac{L_c}{L} - (\frac{k_c}{2L})^2] \tag{2.10}$$

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 the community itself $L_c = 0$ and the sum is negative. Maximum network modularity indicates the best partions. As too many possible partitions exist, we need an algorithmic approach to identify the best separation.

The first algorithm proposed for modularity optimization was **greedy algorithm**. First, it assigns each node to a community and starts with N communities. Then, we should merge each pair of communities and calculate the modularity difference ΔM. We can join those two communities by identifying the pair for which the difference is the largest. It is repeated until all nodes merge into a single community. The best partition is one with the largest M.

Louvain algorithm [68] is an optimization algorithm with better scalability than the greedy algorithm so it can operate on very large networks. Initially, each node is assigned to a different community, and similar to before, we calculate the difference in the modularity moving nodes to one of their neighbouring community. Then we move node i to the community such that modularity becomes larger. It is applied to all nodes until no further improvement can be made. In the second step, we create a weighted network whose nodes are communities identified during the first step. The weight of the links between communities is the sum of the weights between nodes. The number of links inside the community is given as a weighted self-loop. Then, the first and second spets are repeated until there is no more change in the modularity.

2.2.2 Stochastic block model

Another approach for studying the community structure of complex networks, the Stochastic Block Model (SBM), assumes that nodes are clustered in the groups, and the relations between nodes de-

pend on the probabilities for group memberships [77]. In one group, nodes have similar connectivity patterns. To describe the network G with the SBM model, we need to define the following:

- 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_{rs} are the probabilities that edges between groups r and s exist. Note that nodes within one group have the same connection probabilities.

The number of possible nodes between two groups r and s:

$$n_{rs} = \begin{cases} n_r(n_r+1)/2 & \text{if } r = s \\ n_r n_s & \text{if } r = s \\ , \end{cases} \tag{2.11}$$

while the number of possible edges depends on the adjacency matrix A_{ij}:

$$e_{rs} = \frac{1}{1+\delta_{rs}} \sum_{i\in r, j\in s} A_{ij} \tag{2.12}$$

The benefit of this model is that we can **generate** many networks with similar network structure [78]. When model parameters are initialized, the network can be easily generated. For each pair of nodes i and j in network G, we draw a link if random number $r_{ij} < p_{r,s}$.

The likelihood of generating network G for given model parameters is:

$$P(G|p,g) = \prod_{i,j} Pr(i \to j|p,g) = \prod_{(i,j)\in E} Pr(i \to j|p,g) \prod_{(i,j)\notin E} (1 - Pr(i \to j|p,g)) \tag{2.13}$$

In the processes where the connection between two nodes is described with Bernoulli distribution, the likelihood takes the form:

$$P(G|p,g) = \prod_{(i,j)\in E} p_{g_i g_j} \prod_{(i,j)\notin E} (1 - p_{g_i g_j}) \tag{2.14}$$

In the likelihood equation, we iterate over all pairs of nodes, separating the product over edges present in the network and edges that are not present. As all nodes are considered independent, we can switch the product over nodes with the product over groups such that

$$P(G|p,g) = \prod_{(r,s)} p_{rs}^{e_{rs}} (1 - p_{rs})^{n_{rs} - e_{rs}} \tag{2.15}$$

As it is easier to work with the logarithm of the likelihood function, after taking the logarithm of the likelihood function, we get the following expression:

$$L = log(P(G|g,p)) = \sum_{r,s} e_{r,s} ln \frac{e_{rs}}{n_{rs}} + (n_{rs} - e_{rs}) ln(\frac{e_{rs} - e_{rs}}{n_{rs}}) \tag{2.16}$$

Instead of generating networks, the opposite task is network **inference**. For a given network G, and specified the number of communities k, we can use the SBM model to infer the nodes' assignments

into groups, so we need to choose vector g and SBM matrix p such that the likelihood for generating network G is maximized.

The formulation of the SBM model does not consider how to infer the optimal number of groups. Optimizing the likelihood function for different numbers of groups would increase likelihood while each node is not assigned to a different group. In practice, our found community structures for a fixed number of groups, and then the likelihood function could be penalized by the number of model parameters. One approach is calculating the **Minimum description length (MDL)**. For the variable that occurs with probability P(x), the necessary amount of information to describe it is $-log_2 P(x)$. The numerator of posterior probability could be written as

$$P(G|g)P(g) = P(G|p,g)P(p,g) = 2^{-\Sigma} \tag{2.17}$$

where Σ is the data's description length (DL). It means that if we find the network partition that maximizes the posterior distribution, we also find the MDL. The MDL consists of two terms: $\Sigma = -log2(p(G|p,g)) - log2P(p,g)$. In the first part of the equation, the amount of information necessary to describe the model decrease with the number of groups. The second contribution comes only from the model, and as the model becomes more complex, with a larger number of groups, this part increases. The optimal solution represents the balance between these two terms in the MDL equation.

This SBM model has many variants motivated by specific properties of real data. For example, for degree heterogeneous networks, there is degree corrected SBM [73]. In some social networks, users can belong to more than one group, which can be modelled with mixed membership SBM. Other extensions include application to bipartite, weighted network, and hierarchical model [79]. Many community detection algorithms define the community as an assortative structure. With the SBM model, such limitations do not exist, and it is possible to directly use statistical inference for discovering core-periphery structures or even networks with bipartite structures.

2.2.3 Core-periphery structure

Core-periphery structure describes a network whose nodes are divided into two communities, densely connected core and less connected periphery [69]. With the average probabilities of edges within each group as p_{11} and p_{22}, and between groups p_{12}, the core-periphery structure is defined under the condition $p_{11} > p_{12} > p_{22}$. The simple method for finding core-periphery structures assumes that core nodes have higher degrees in the core than in the periphery. Another simple method is to construct k-cores [70]. K core is a group of nodes connected to at least k other members. K-cores form a nested set and become denser with higher k. The core-periphery structure can be detected by optimizing the measure similar to modularity, as defined by Borgatti and Everett [71]. Their goal is to find the division that minimizes the number of edges in the periphery. So they define the score function as equal to the number of edges in the periphery minus the expected number of such edges placed randomly. $\rho = \frac{1}{2}\sum_{ij}(A_{ij} - p)g_i g_j$.

Another way to detect core-periphery structure is to use the inference method based on fits to a Stochastic Block Model (SBM) [72, 73]. In this method, we fit the observed network to a block model with two groups, such that edge probabilities have the form $p_{11} > p_{12} > p_{22}$. Vector $\theta_i = r$ indicates that node i is in block r, while SBM matrix $\{p\}_{2x2}$, 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 [55]. Probability of having model parameters θ, p given network G is proportional to the likelihood of generating network G, prior of SBM matrix $P(p)$ and prior on block assignments $P(\theta)$: $P(\theta, p|G) = P(G|\theta, p)P(p)P(\theta)$, while the likelihood function takes following form: $P(G|\theta, p) = \prod_{i<j} p_{r_i s_j}^{A_{ij}}(1 - p_{r_i s_j})^{1-A_{ij}}$, where A_{ij} is a number of edges between nodes i and j. The prior $P(p)$ is modified for core-periphery model such that $P(p) \sim I_{0<p_{22}<p_{12}<p_{11}<1}$, while prior

$P(\theta)$ consists of three parts: probability of having 2 blocks; given the number of layers probability $P(n|2)$ of having groups of sizes n_1, n_2 and the probability $P(\theta|n)$ of having particular assignments of nodes to blocks.

2.3 Network models

2.3.1 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 $r \leq p$

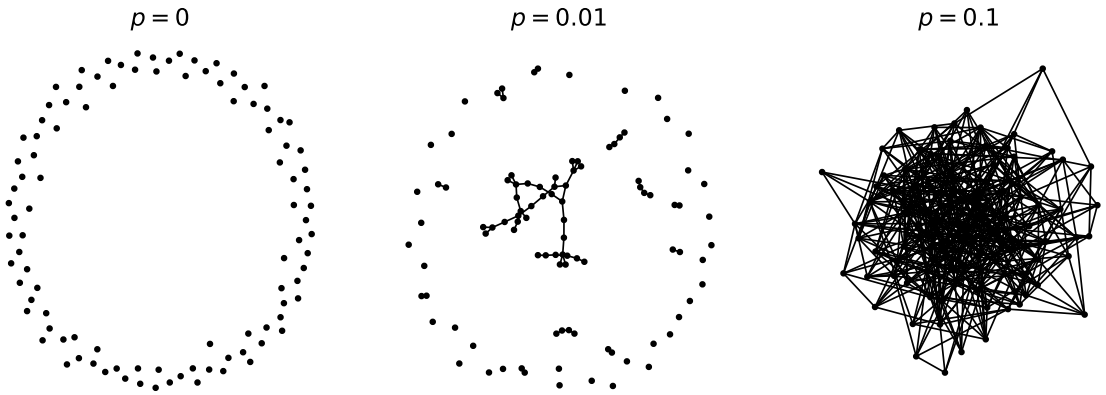

Figure 2.2: Erdős-Rényi graph with $N = 100$ nodes and different linking probabilities p.

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 of 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 a complete graph with N nodes, the number of edges is $N(N-1)/2$. As the probability of drawing every edge is p, the **average number of links** is given as

$$\langle L \rangle = \frac{N(N-1)}{2} p \tag{2.18}$$

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

$$\langle k \rangle = (N-1)p \tag{2.19}$$

The **degree distribution** of ER random graph follows the binomial distribution [42].

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k} \tag{2.20}$$

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 $\langle k \rangle$

$$p(k) = \frac{1}{k!} e^{-\langle k \rangle} \langle k \rangle^k \tag{2.21}$$

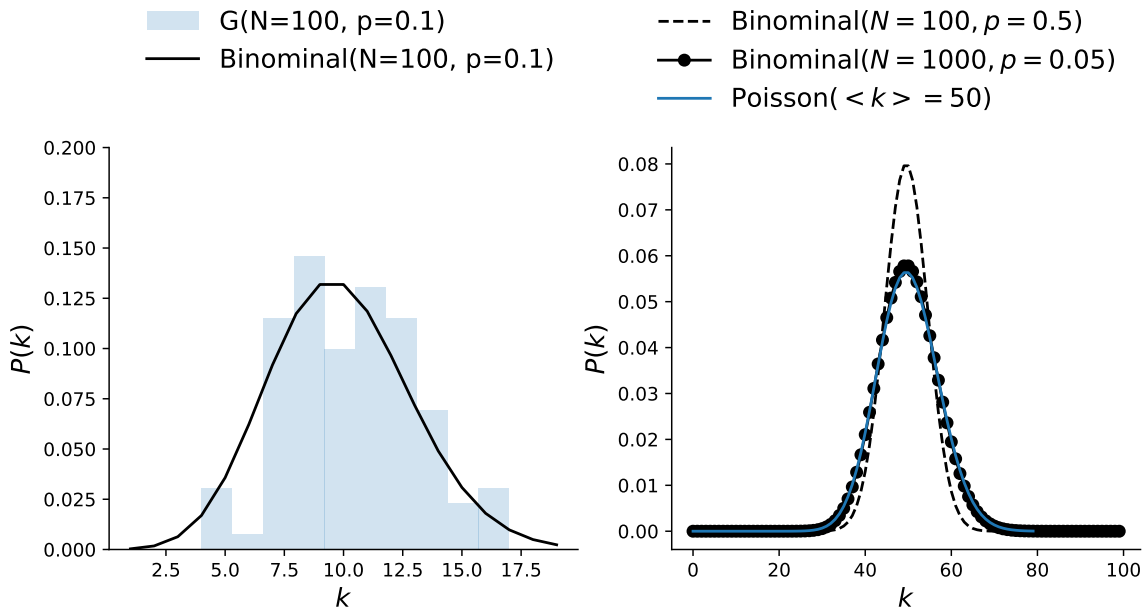

Figure 2.3: Degree distribution of ER graph. The degree distribution of small networks follows binominal. Larger networks are better approximated with Poison distribution, and degree distribution for fixed average degree $< k >$ becomes independent of the network size.

The random graph has a very small **average path length**, it is given as $\langle l \rangle = \frac{lnN}{ln\langle pN \rangle}$ that is characteristic of many large networks [74]. The clustering coefficient is proportional to linking probability, $\langle C \rangle = p$, so we find a small clustering coefficient in large random networks, contrary to real-world networks.

Figure 2.2 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 whose 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 the average degree is $\langle k \rangle = 1$, which gives us: $p_c = \frac{1}{N-1}$, meaning that all nodes have degree larger than one [42]. When the $\langle k \rangle < 1$, the network is in the subcritical 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.

2.3.2 Small-world networks

Inspired by the idea that real-world networks are highly clustered, and the average distance is small, Watts and Strogatz [26] 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.

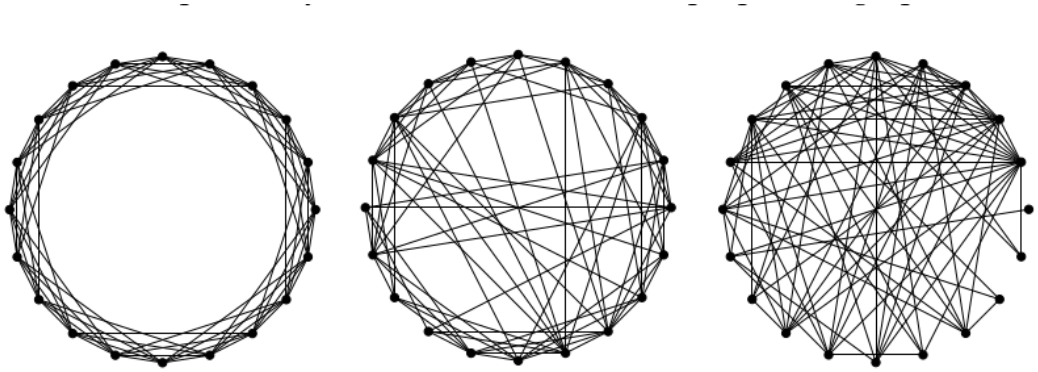

Figure 2.4: Watts and Strogatz graph model creation.

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

2.3.3 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. Growth means that at each time step, new nodes are added to the network. The simplest model that produces scale-free networks is the Barabasi-Albert model [27].

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

The preferential attachment is important for generating a system with scale-free properties. In the real system, the linking between nodes is not a random process; the preference for specific types

of nodes exists. For example, popular web pages can quickly get more visits, or it is expected 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:

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j} \tag{2.22}$$

As at each step one node arrives, 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 degrees in time.

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j} = m\frac{k_i}{m_0 + 2mt} \tag{2.23}$$

Note that the 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 $t > t_i$, it has a degree that grows as the square root of time; it also shows that younger nodes easily acquire a larger degree.

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} \tag{2.24}$$

Degree distribution follows power-law, and for large k is approximated with $P(k) = k^{-\gamma}$, such that $\gamma = 3$. More precisely, the degree distribution has form [75]:

$$P(k) = \frac{2m(m+1)}{k(k+1)(k+2)} \tag{2.25}$$

For large k, it is exactly the power law. It is also independent of the time and size of the system, meaning the emergence of a 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 rescaling $p(k)/m^2$, they fall on the same line [42].

As the network grows, nodes with larger degrees become bigger, so we end up with few nodes with many links, called hubs. The **network diameter**, represents the maximum distance in network, $d \sim \frac{lnN}{lnlnN}$ [74]. The diameter grows slower than lnN, making the distances in the BA model smaller than in the random graph. The difference is found for large N. It is known that the BA network has hubs that shorten the path between less connected nodes. Also, if hubs are removed from the network, the network easily partitions into several components, losing its properties. The **clustering coefficient** of the BA model follows $C \sim \frac{lnN^2}{N}$ [74]. It differs from clustering found in random networks, and BA networks are generally more clustered.

The combination of the growth and preferential attachment linking is crucial for getting scale-free networks [27]. For example, eliminating the preferential attachment; in a growing network with random linking, degree distribution is stationary but follows exponential. In contrast, the absence of growth leads to the non-stationary degree distribution. When a number of nodes is fixed, the network grows only in the number of links, such that randomly chosen node i connects to node j according to probability Π. In the beginning, the degree distribution follows the power law, the same as in the BA model. As more links are added to the network, the distribution changes its shape; first, the peak appears, while at the end network becomes a complete graph, where all nodes have the same degree.

2.3.4 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 [76]:

$$\Pi(k_i) = {k_i}^{\beta} \tag{2.26}$$

The probability that a 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 the BA model, where degree distribution follows the power law. When $\beta = 0$, linking probability becomes uniform; i.e. it corresponds to a random network model, and the degree distribution is Poisson; there is exponential decay.

For $\beta > 1$, preferential attachment effects are increased, leading to super hubs' emergence. The hub-and-spoke network appears in this regime, where almost all nodes are connected to a few high-degree nodes [76].

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

2.3.5 Aging 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 [28].

$$\Pi_i(t) \sim k_i \tau_i^{\alpha} \tag{2.27}$$

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.

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 [51]

$$\Pi_i(t) \sim k_i(t)^{\beta} \tau_i^{\alpha} \tag{2.28}$$

As before, depending on model parameters network evolves into different structures:

- 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.

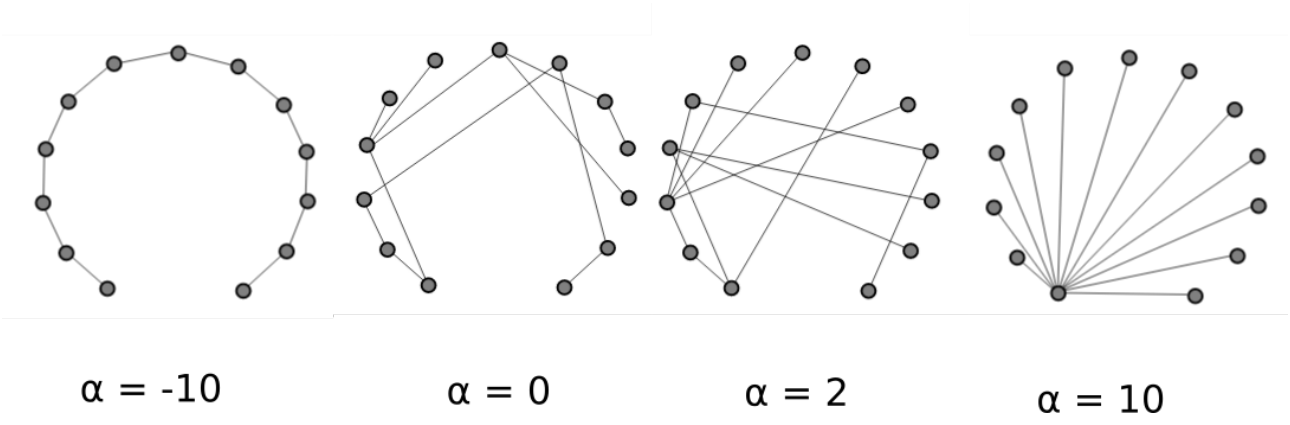

Figure 2.5: Aging model

- Scale-free property can be produced along the critical line $\beta(\alpha^*)$ in the $\alpha - \beta$ phase diagram, see Figure 2.6.
- 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

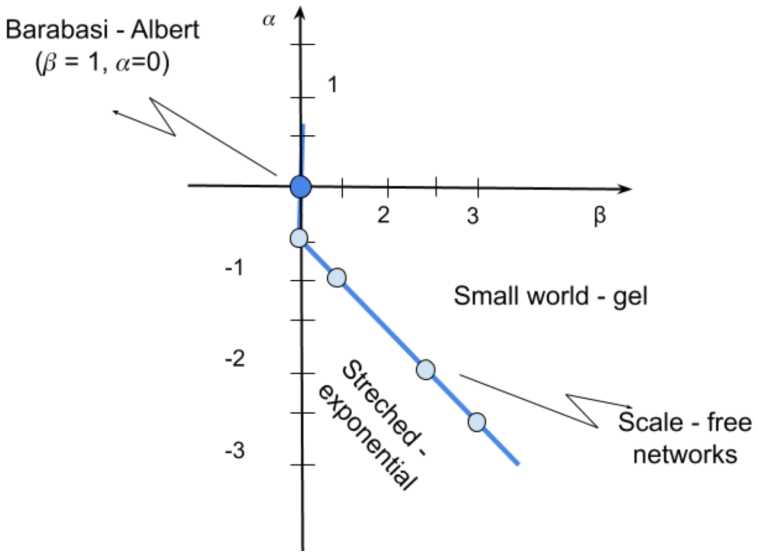

Figure 2.6: Phase diagram of aging network model

2.4 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 randomly, and any two nodes are linked with the same probability p, we expect the binomial distribution. 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 [42].

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 [42]. 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 lognormal 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.

2.4.1 The properties of distributions

Power-law distribution. The power-law distribution [80, 81] is defined as

$$p(k) = Ck^{-\gamma} \tag{2.29}$$

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

The distribution can take discrete and continuous values, 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 [41]. We'll find that these criteria are not satisfied by any other distribution.

$$\frac{p(k)}{p(2k)} = \frac{Ak^{-\gamma}}{A(2k)^{-\gamma}} = 2^{\gamma} \tag{2.30}$$

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 $\gamma = -p(1)/p^{'}(1)$ leads us to the conclusion that if the 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 [82].

$$f(y) = \frac{1}{2\pi\sigma}e^{-(y-\mu)^2/2\sigma^2} \tag{2.31}$$

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

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-(log(x)-\mu)^2/2\sigma^2} \tag{2.32}$$

The lognormal distribution has finite mean $e^{\mu+1/2\sigma^2}$, and the variance $e^{2\mu+\sigma^2}(e^{\sigma^2-1})$. [80]. Despite the finite moments, the lognormal 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 lognormal distribution [41, 80]. The lognormal 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,

$$S_t = \epsilon_t S_{t-1} \tag{2.33}$$

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.:

$$S_t = \epsilon_t S_{t-1} = \epsilon_t \epsilon_{t-1} ... \epsilon_2 \epsilon_1 S_0 \tag{2.34}$$

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

$$ln(S_t) = ln(S_0) + \sum_{i=0}^{t} ln(\epsilon_i) \tag{2.35}$$

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 lognormal distribution.

The multiplicative processes generate the lognormal distribution. Introducing a threshold in the multiplicative process leads to the power law. For example, in the Champernowne model [41], 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 a higher class. In this process, the distribution of incomes emerges as the power-law distribution.

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

$$p(k) = Ck^{-\gamma} e^{-\lambda k} \tag{2.36}$$

where $k > 0$ and $\gamma > 0$. This function combines the power-law, and exponential terms responsible for an exponentially bounded tail [42]. Taking the logarithm $ln(p(k)) = lnC - \gamma lnk - \lambda k$, when $k << 1/\lambda$ the second term dominates, so distribution follows the 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:

$$p(k) = ck^{\beta-1} e^{-(\lambda k)^{\beta}} \tag{2.37}$$

the parameter β is stretching exponent determining the properties of the function $p(k)$ [42]. 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.

2.4.2 Plotting the distributions

The first step in analyzing 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 [41]. If data are from power law distribution on the double logarithmic scale, they will look linear:

$$logp(k) = \gamma log(k) + c \tag{2.38}$$

On the log-log scale, we notice that noise in the tail of the distribution data exists. 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 [83, 41, 84]. 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 the number of nodes found in the bin b_n and $k_n = \sum_i k_i/N_n$ is the 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 the 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 2.7 shows how different distributions look 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 a 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.

2.4.3 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 maximize the likelihood of the data to find the distribution parameters. For a 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 [84]. The likelihood function tells us how likely it is to have the given data if the distribution parameters are θ.

$$L(\theta|x_1, ...x_n) = \prod_{i=1}^{i=n} p(x_i|\theta) \tag{2.39}$$

The parameter that maximizes 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 optimization for distributions where an analytical solution is unavailable. In practice is much easier to work with the logarithm of the likelihood function, $log(L) = \sum_{i=1}^{i=N} p(\theta|x_i)$, because then 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 a discrete distribution, the solution may be obtained by 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 [85] 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

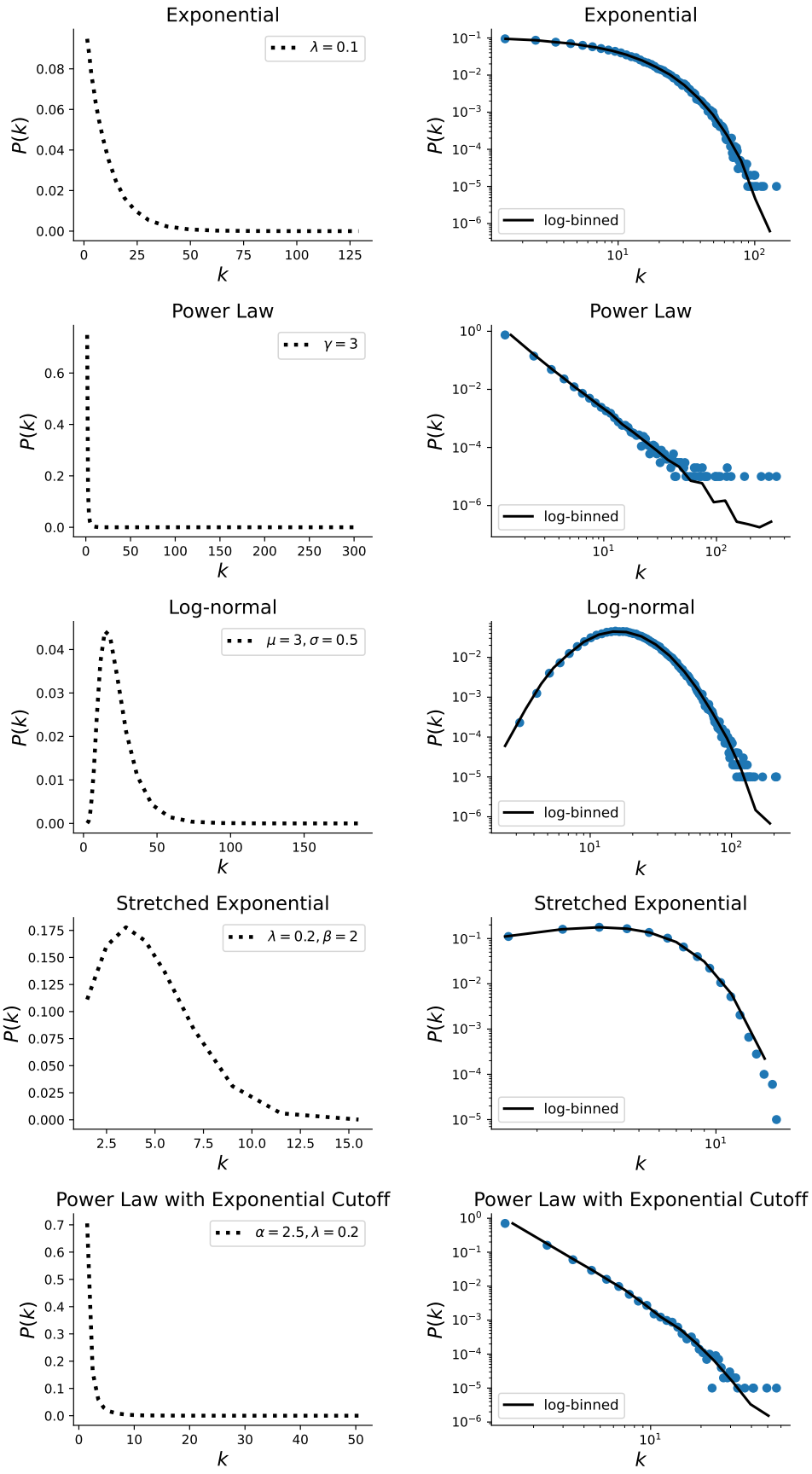

Figure 2.7: Probability distributions on a linear and double logarithmic scale.

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, $D = max|S(x) - P(x)|$.

First, we fit empirical data to get model parameters and calculate the KS statistics of this fit [85]. Then, many synthetic data sets are generated with model-optimized 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 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 [85]. 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

$$\mathcal{R} = \sum_{i=1}^{n} [logp_1(x_i) - logp_2(x_i)] \tag{2.40}$$

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 the data, and then $R < 0$, 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

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}} \int_{-\infty}^{-|\mathcal{R}|} e^{-x^2/2n\sigma^2} dx + \int_{|\mathcal{R}|}^{\infty} e^{-x^2/2n\sigma^2} dx \tag{2.41}$$

Here we use the standard two-tail hypothesis test [85], 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 $R = 0$.

2.5 Fractal analysis

One of the approaches in studying complex systems is detecting the time series of selected variables [86]. In complex systems, the periodic behaviour of time series is not limited to one or two characteristic frequencies. They extend over a broad spectrum and fluctuations on many time scales and broad distributions [87, 88]. In these cases, the system's dynamics are characterized by scaling laws, 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, characterizes the type of self-affinity.

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

.

2.5.1 Long and Short-term correlations

The time series are persistent, meaning that large values usually follow a large value [86]. 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, 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.

$$C(s) = \langle \Delta x_i \Delta x_{i+s} \rangle = \frac{1}{N-s} \sum_{i=1}^{N-s} \Delta x_i \Delta x_{i+s} \tag{2.42}$$

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 characterized by Hurst exponent $H = 1 - \gamma/2$.

A direct calculation of the $C(s)$ is complex 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.

2.5.2 Rescaled range analysis

Hurst proposed a method called the **rescaled range analysis** R/S, [89]. 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 [86].

2.5.3 Fluctuation analysis

The **fluctuation analysis** is based on the random walk theory [86]. For given time series $\{x_i\}$ with length N, we first define the global profile in the form of cumulative sum, equation 2.43, where $\langle x\rangle$ represents the average of the time series.

$$Y(j) = \sum_{i=0}^{j}(x_i - \langle x\rangle), \quad j = 1, ..., N \tag{2.43}$$

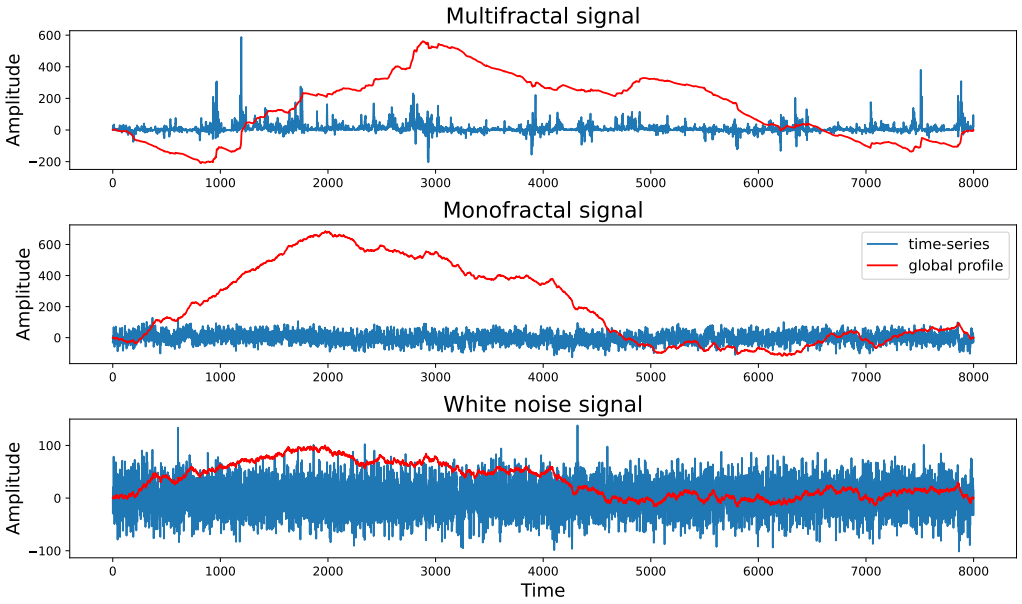

Figure 2.8: Multifractal, monofractal and white noise signals.

The profile of the signal Y is divided into $N_s = int(N/s)$ non-overlapping segments of length s. The last segment will be shorter if N is not divisible with s. 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.

$$F_2(s) = [\frac{1}{2N_s} \sum F^2(\nu, s)]^{1/2} \sim s^H \tag{2.44}$$

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 the 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 do not have to be linear [90]. 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) [91]. 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:

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \left[Y(j) - p^m_{\nu,s}(j)\right]^2 \tag{2.45}$$

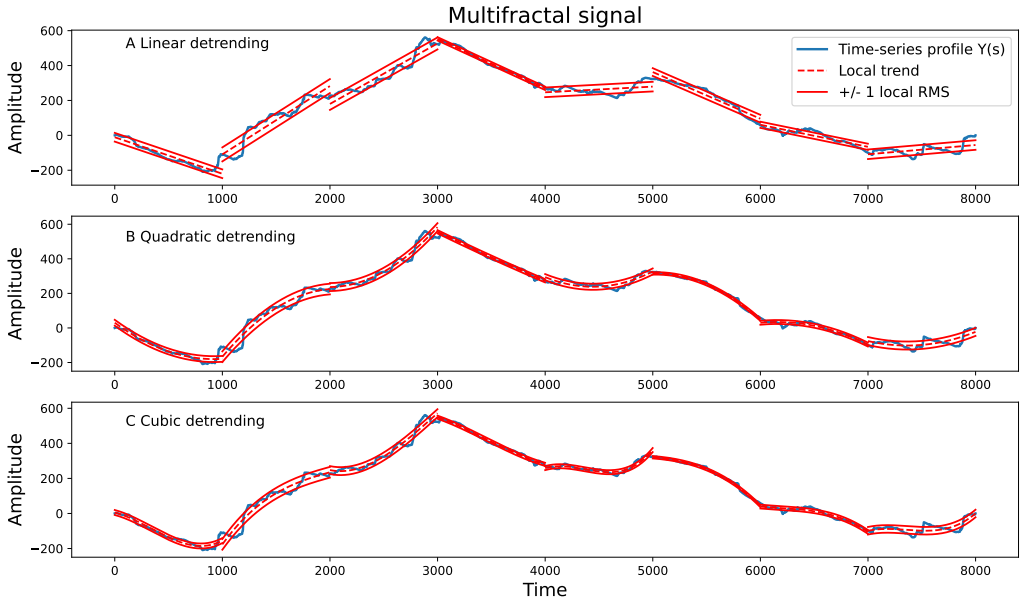

Figure 2.9: Detrending the signal for the segments of length $s = 1000$.

2.5.4 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 [92, 93] 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 to the DFA method. The value of $H(0)$, which corresponds to the limit $H(q), q - 0$, cannot be determined directly because the exponent diverges. Instead, the logarithmic averaging procedure has to be considered.

$$F_0(s) = \exp \left\{ \frac{1}{4N_s} \sum_{\nu}^{2N_s} ln \left[F^2(\nu, s) \right] \right\}, q = 0 \tag{2.46}$$

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 $H(q)$, see Figure 2.10

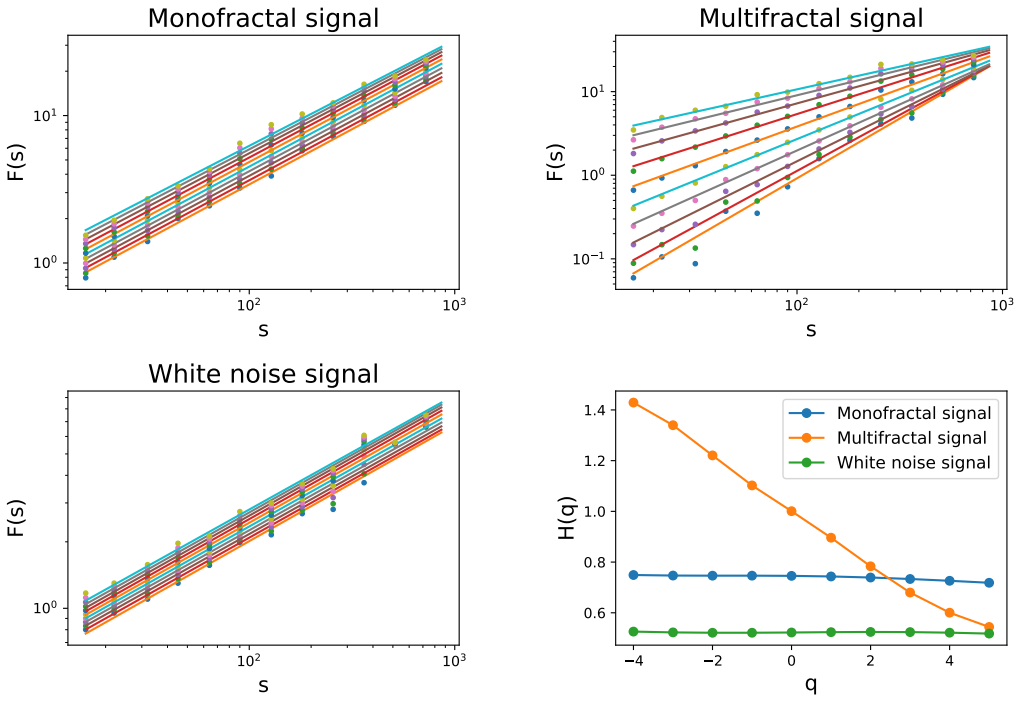

Figure 2.10: Fluctuating function and Hurst exponent.

Dependence of the fluctuating functions on the scale for monofractal, multifractal and white noise signals, and the dependence of the Hurst exponent H on the scale 1 q for different types of signal (bottom right).

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 depend 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.

2.6 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 dynamic property of the trust. Furthermore, the important property of the 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) [56], we can quantify the user reputation R_n after each interaction using equation 2.47, where n is the number of interaction $n \in 1, N$.

$$R_n = R_{n-1}\beta^{\Delta_n} + I_n \tag{2.47}$$

The first part of the equation considers the reputation value after the 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}:

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a} \tag{2.48}$$

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.

$$I_n = I_{b_n}(1 + \alpha(1 - \frac{1}{A_n + 1})) \tag{2.49}$$

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 $A_n = 1$.

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 α and β values lead to faster reputation decline, as shown in Figure 2.11 - left panel. 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 inactive for a larger period. The parameter α is the most important influence on burst behaviour, where larger α leads to higher reputation values.

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 $\Delta_d = (t_d - t_{n-1})/t_a$.

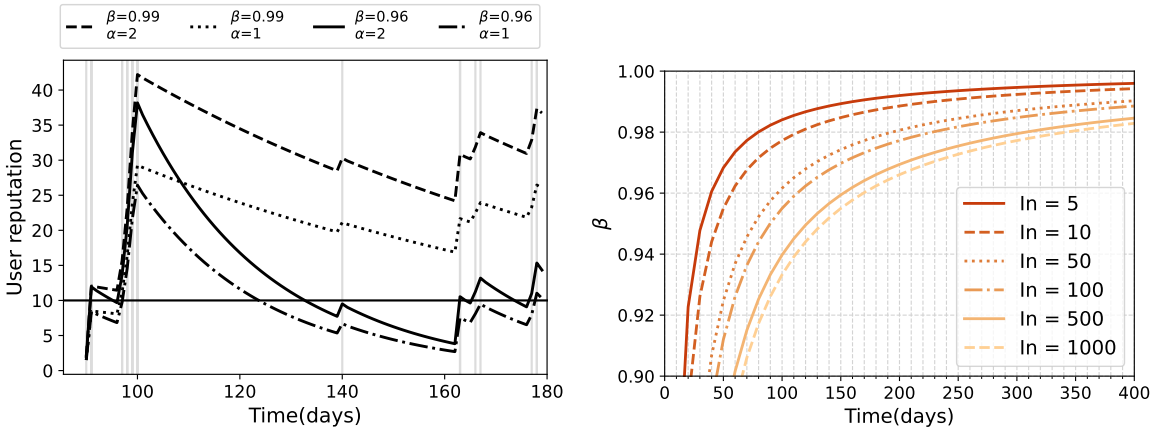

Figure 2.11: User reputation.

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 in Figure 2.11, - right panel, the parameter $t_a = 1$day, while we vary different starting reputation levels I_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, reputation with starting value 5 needs around one month to decay below the threshold. For higher reputations, 500 or 1000, the decay period is around 5 months.

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 system, we can also fit the model parameters to mimic the actual reputation dynamic.

Chapter 3

Driving signals

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. 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 nonlinear 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 complex networks. Differences between networks with the same number of nodes and links can be observed by analyzing connectivity patterns. Figure 3.1 summarizes our goals.

3.1 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:

$$\partial_t N_k = \sum_{j=1}^{M(t)} r_{k-j \longrightarrow k} N_{k-j} - \sum_{j=1}^{M(t)} r_{k \longrightarrow k+j} N_k + M(t)\delta_{k,L}. \tag{3.1}$$

We add $M(t)$ nodes with L links at each time step. As multiple 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. The first term in the equation describes nodes with degree $k \in \{k-M(t), \ldots, k-1\}$ that getting degree k,

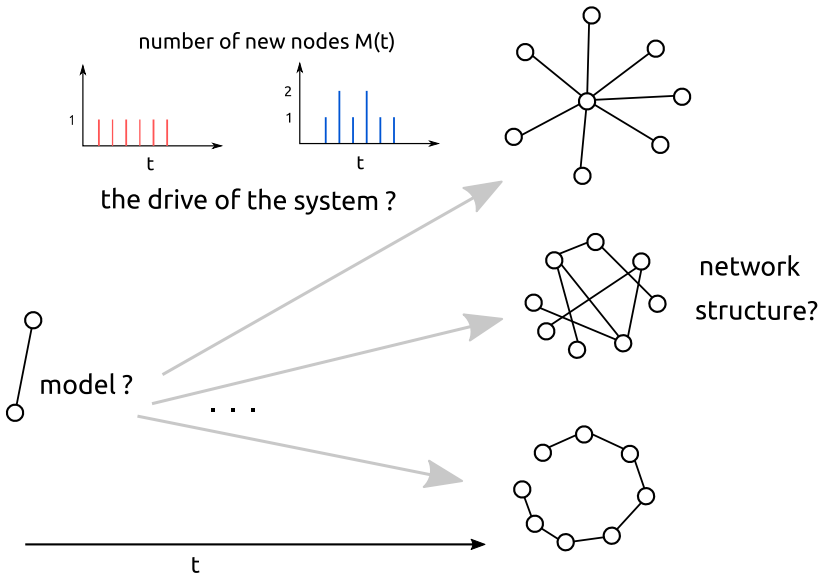

Figure 3.1: The open question is how nonlinear signals, in combination with the network model, influence the network's structure. Under what circumstances do networks have the scale-free, hub-spoke or chain structure?

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 the model, we choose the 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 [29]. 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 $L > 1$; 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 a 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 the new nodes' time series. We focus on the growth of two real systems, the **TECH** [94] community in the Meetup website and on two months of **MySpace** [95] social network.

3.1.1 Time-series from real systems

MySpace signal is the number of new members who appear for the first time in the data. 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, 3.2. 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 3.2. We can easily destroy trends and cycles if we randomize the MySpace signal. 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, the randomized MySpace signal no longer has long-range correlations; the Hurst

exponent indicates short-range correlations $H = 0.5$, and the signal becomes monofractal.

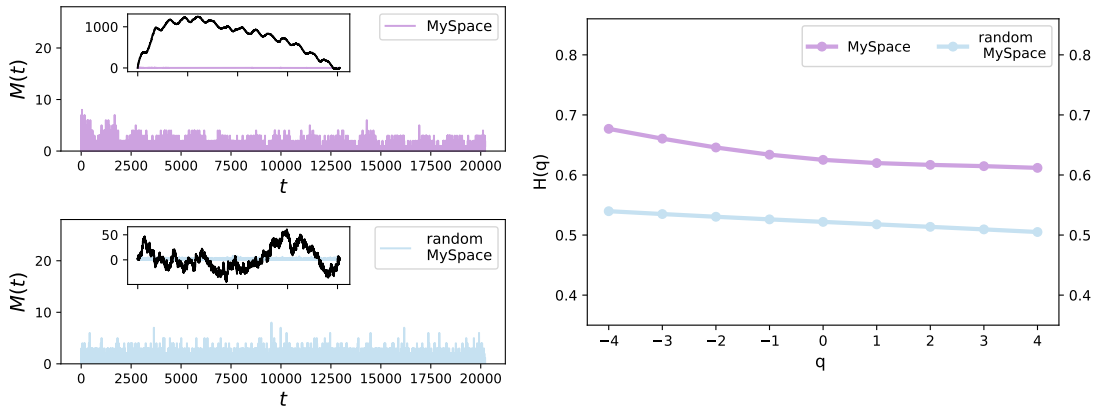

Figure 3.2: MySpace signal, the random MySpace signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q. (right pane)

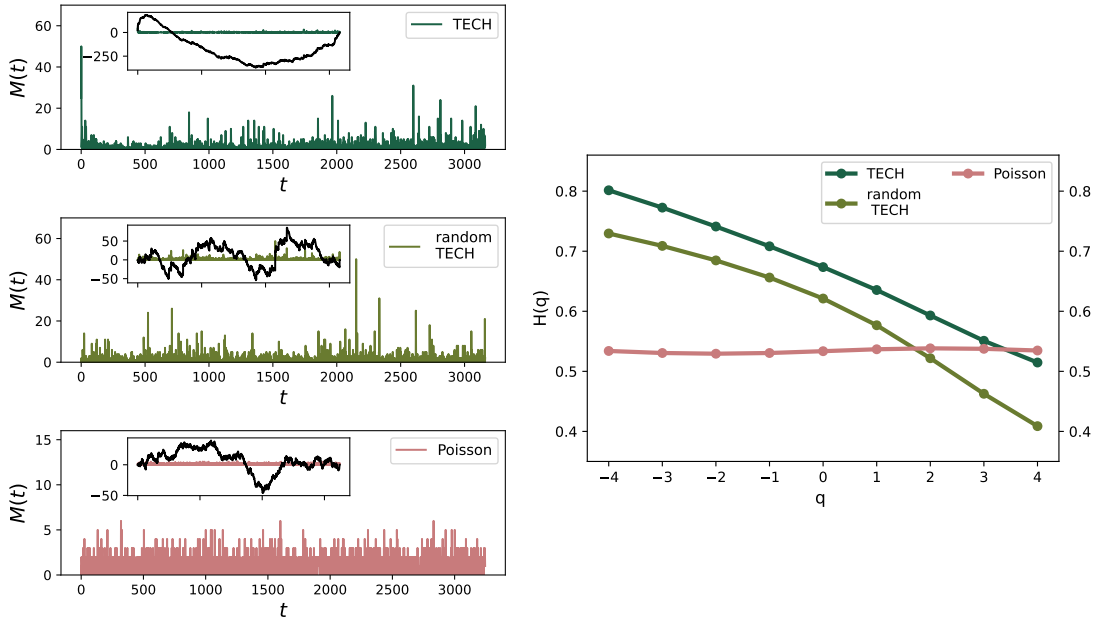

Figure 3.3: TECH signal, the random TECH signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q. (right pane)

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 events. 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, Figure 3.3. 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 can easily destroy trends and cycles, but the signal keeps multifractal properties, meaning that broad probability distribution can not be eliminated. Therefore, we generate the uncorrelated signal from the Poissonian probability distribution. The length of this signal is $T = 3246$, while we keep the number of nodes N the same as in the TECH signal.

3.1.2 D-measure

We can compare the networks with the same number of nodes and links generated with growth signals with different properties. We use a 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 [63], 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 average 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 [63].

Figure 3.4 presents the results for D-measure. The most significant distance between networks is along the critical line $\beta(\alpha^*)$ of the aging model. The fluctuations present in the signal mainly influence the scale-free networks. Structural differences exist for networks away from this line, but they are much smaller. The D-measure is close to zero for gel small-world networks, $\beta > \beta^*$. Under critical line, $\beta < \beta^*$, the D-measure depends on the properties of the signal. If we fix network density L, the position of the critical line is independent of the properties of the signal. Still, with higher link density, the critical line slightly moves toward larger β; see Figure 3.4.

In the region around the critical line, we find that the 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. The difference exists for networks generated with uncorrelated signals: random MySpace and Poisson, 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. Hubs emerge faster in signals, such as TECH, where there are peaks in the number of new users. 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.

3.1.3 The structure of networks

We examine degree distribution, degree correlations and clustering coefficient of networks generated by real signals. These measures have provided 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 3.5 shows properties of networks generated with model parameters $L = 2$, $\alpha = -1.0$, $\beta = 1.5$, that lie on the 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 signal. Networks generated with real signals average neighbouring degree $\langle k \rangle_{nn}(k)$ and clustering coefficient $c(k)$ depend on node degree. In contrast, networks generated with constant and randomized signals weakly depend on the degree k.

We also find structural differences between networks, obtained with model parameters under the

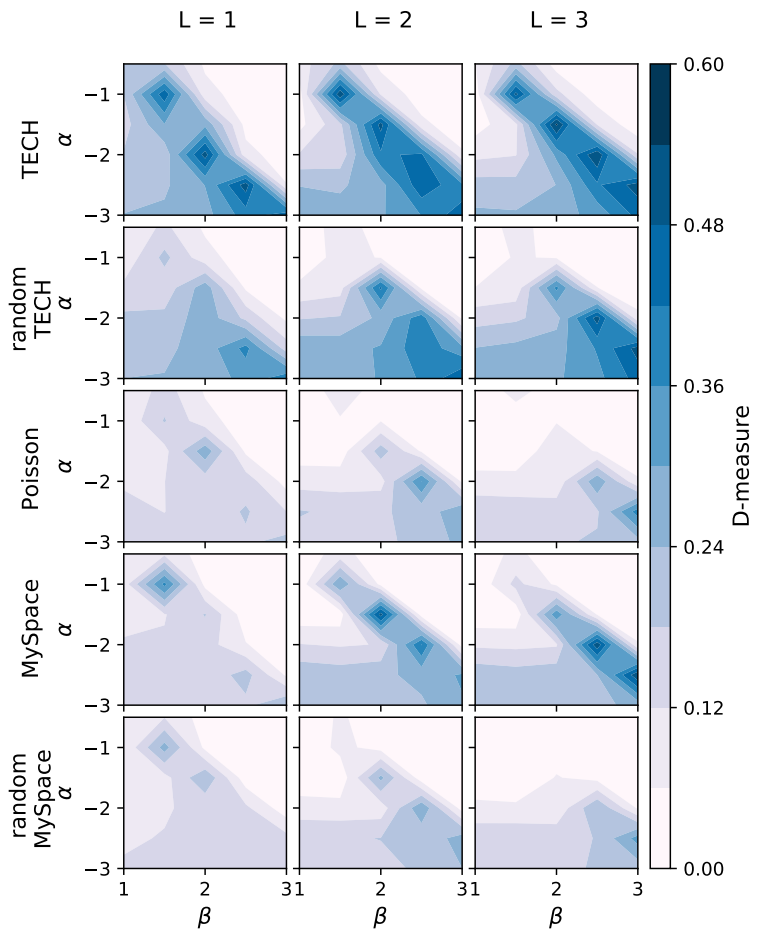

Figure 3.4: The comparison of networks grown with growth signals shown in figures 3.3 and 3.2 versus ones grown with constant signal $M = 1$, for the value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in each time step. The compared networks are of the same size.

critical line $\alpha < \alpha^*$, see Figure 3.5. The difference is mainly found in the 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 3.5. 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.

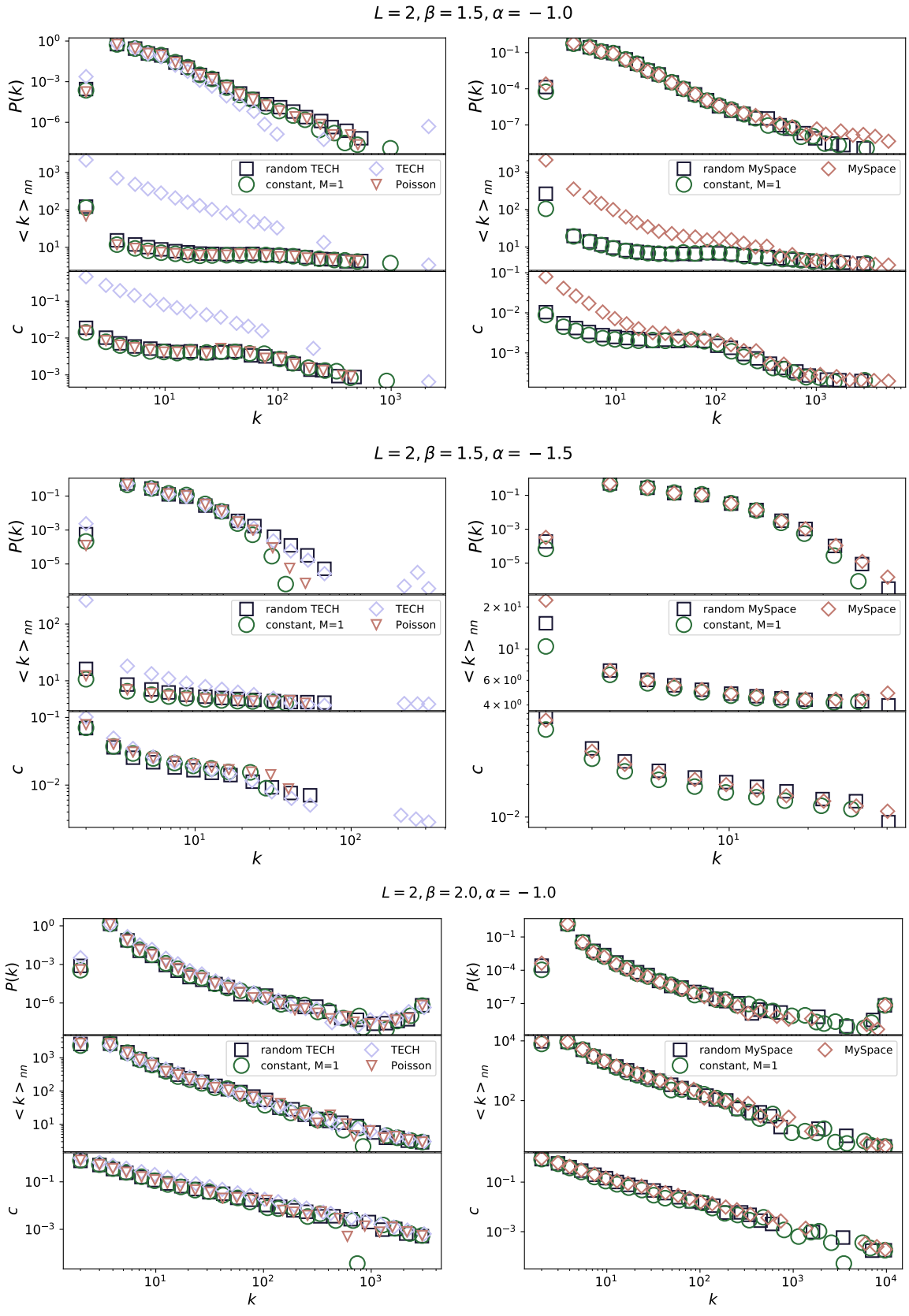

Figure 3.5: Degree distribution, the dependence of average first neighbour degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $\alpha = -1.0$, $\beta = 1.5$ and $L = 2$ for all networks. The networks are from the 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.

3.2 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 analyze their influence separately. 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 in time series using Fourier filtering transform method [96].

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 $S(f)$, which is a function of the frequency f. For the long-range correlated data, it depends on coefficient $\beta = 1 - \gamma$ and has the form:

$$S(f) \sim f^{-\beta} \tag{3.2}$$

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 the power spectrum expected for data with long-range correlations.
- calculate the inverse Fourier transform x_i. It converts data to the time domain where the 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 different values of the Hurst exponent; see Figure 3.6. The obtained signals are round to integers, and the mean values of signals are close to 4.

As before, we focus on the region of the model phase diagram with negative α and positive β as the transition line from stretched-exponential across scale-free to the small world-gel networks are found. We take a range of parameters $-3 \leq \alpha \leq -0.5$ and $1 \leq \beta \leq 3$ with steps 0.5, and we also vary 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 the 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 Figure 3.7, are obtained by averaging the D-measure between all possible pairs of generated networks.

The higher values of the D-measure are found in the region of critical line $\beta(\alpha^*)$. The most considerable influence is on networks with the scale-free distribution. Comparing D-distance in only one point of the phase diagram, for example, $L = 1, \alpha = -2.5, \beta = 2.5$, we find that when the Hurst exponent is more prominent, correlations in the signal make a bigger impact on the network structure. D-measure between networks grown by signal with Hurst exponent $H = 1.0$ and the constant signal is $D(H = 1.0, M = 4) = 0.405$, while between networks grown with a signal with $H = 0.8$ and the 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.

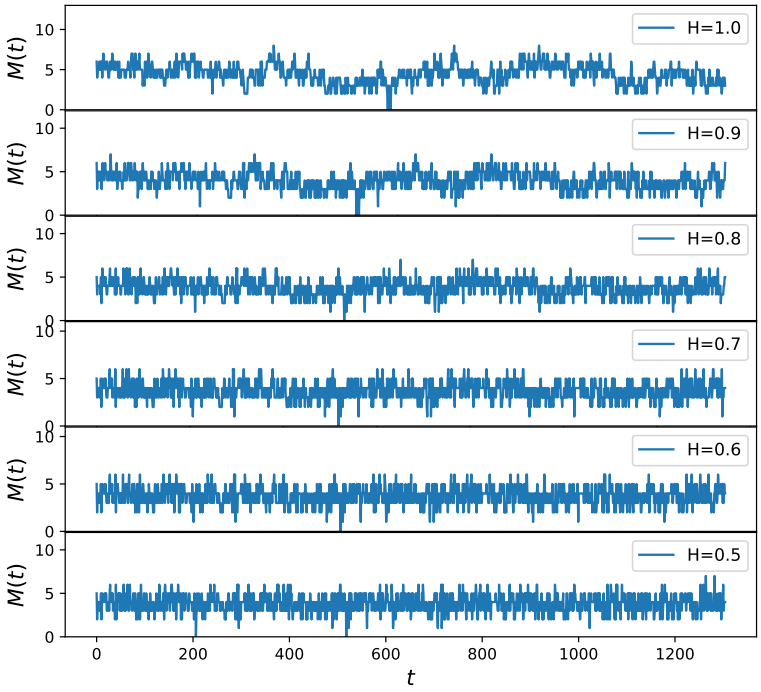

Figure 3.6: Monofractal signals generated with Fourier filtering method for different Hurst exponents

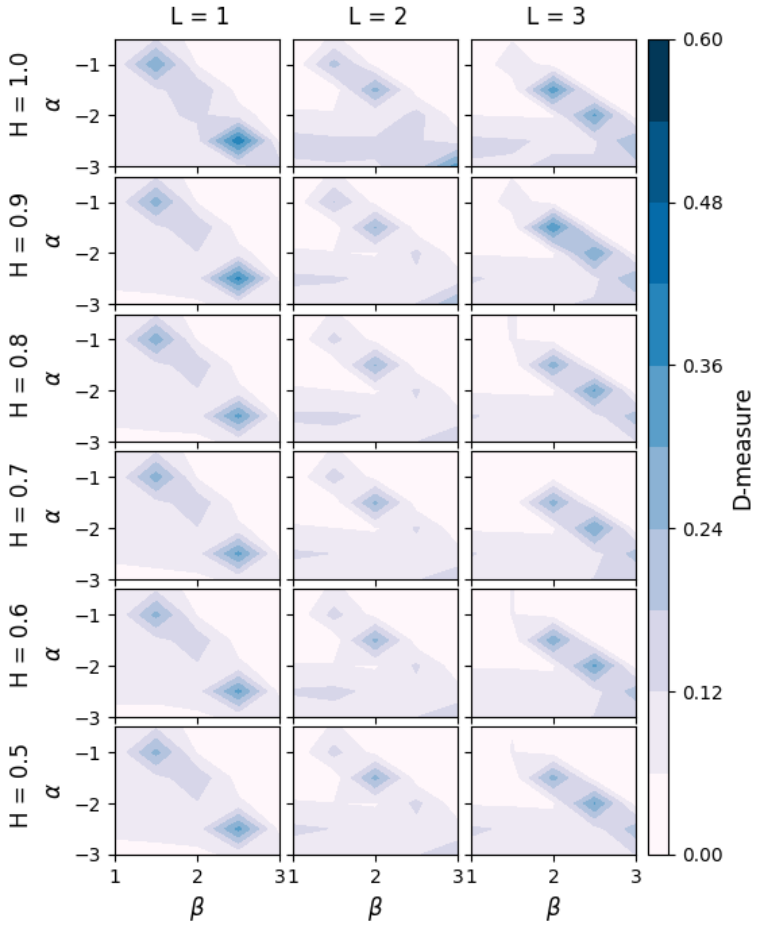

Figure 3.7: D-distance between networks generated with different long-range correlated signals with a fixed value of Hurst exponent and networks generated with constant signal M=4.

We further explore the assortativity index and clustering coefficient of generated networks. Figure 3.8 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 parameters below critical line values, $\alpha = -2.5, \beta = 1.5$ r does not depend on the Hurst exponent. Above the critical line are small-world networks, and 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 grows slightly with link density.

In the region of critical parameters, the assortativity index depends on the value of the Hurst exponent. Signals With Hurst exponent $H > 0.8$ have a larger influence on the assortativity index. Networks become more disassortative; see the line for parameters $L = 1, \alpha = -2.5, \beta = 2.5$ in Figure 3.8. The long-range correlations have a stronger effect on the evolution of networks with lower density.

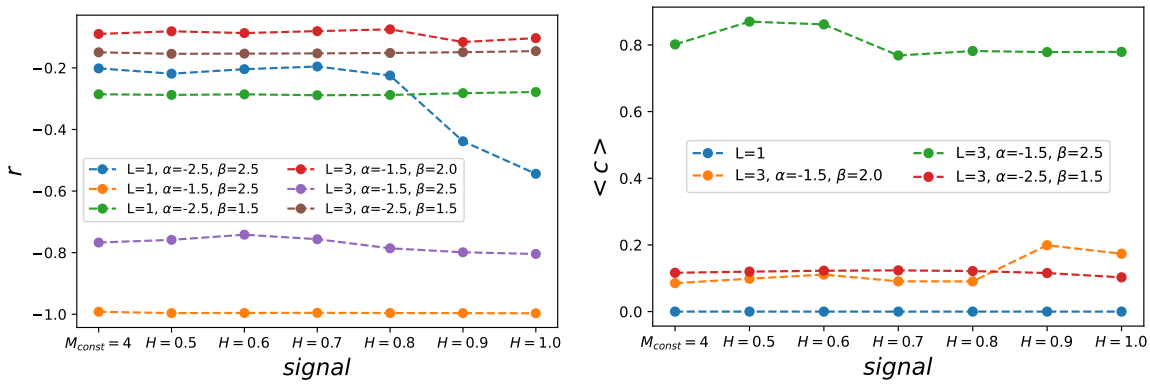

Figure 3.8: Mean assortativity index for networks generated with different model parameters α, β, L and different long-range correlated signals with Hurst exponent H.

Figure 3.8 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 $L = 3, \alpha = -2.5, \beta = 1.5$, the clustering coefficient is constant and low. Similar values are obtained for the clustering coefficient for critical parameters $L = 3, \alpha = -1.5, \beta = 2.0$, but for Hurst exponent $H > 0.8$ clustering coefficient increases. 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.

3.3 Conclusions

In this chapter, we focused on the properties of growth signals and their influence on the system. The network grows at a constant rate in the simplest complex network models. In reality, growth signals are not constant, they are temporally correlated, and the main question is what impact they have on the complex networks. We combined the ageing model with nonlinear growth while we used real and computer-generated long-range correlated signals for growing signals. The network structure depends on the type of signals.

The ageing model can generate different complex networks depending on the model parameters. Our results showed that the most significant difference between networks generated with a constant and fluctuating signal is found on the critical line, where networks have broad degree distribution.

While temporal correlations do not affect the degree distribution, the networks generated with fluctuating signals are more clustered and have more significant degree-degree correlations. The D-measure indicates that structural differences exist even for networks generated with white noise. For multifractal signals, we find the larger values of the D-measure. Furthermore, if we focus only on monofractal signals, characterized by the fixed value of Hurst exponent, H, the difference between networks rises with H.

Away from the critical line, the fluctuations do not strongly influence the network structure; D-measure is close to zero. In small-world networks, super-hubs emerge, and no matter how strong correlations, trends or cycles exist in the signal, the structure of small-world networks does not change. Similar conclusions are found under the critical line, where networks with stretched exponential degree distribution appear. As $\alpha << \alpha^*$, the new nodes attach to close ancestors, and monofractals do not impact the network structure. Only signals with multifractal properties may contribute to the formation of hubs, which is reflected in larger D-measure between networks.

Previous research on temporal networks [43] has shown that edge activation properties impact the complex system's dynamics. Also, different studies indicated the importance of fluctuating signals. Our results imply that modelling the social and technological networks should include non-constant growth. In combination with local linking rules, the properties of growth signals can significantly alter the network structure.

Chapter 4

The growth of social groups

4.1 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 we downloaded data using the Meetup API. 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 the 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, and we 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.

4.1.1 The empirical analysis of social groups

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

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

joined group g. For each 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 over 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 $R = log(S_i(t)/S_i(t-1))$.

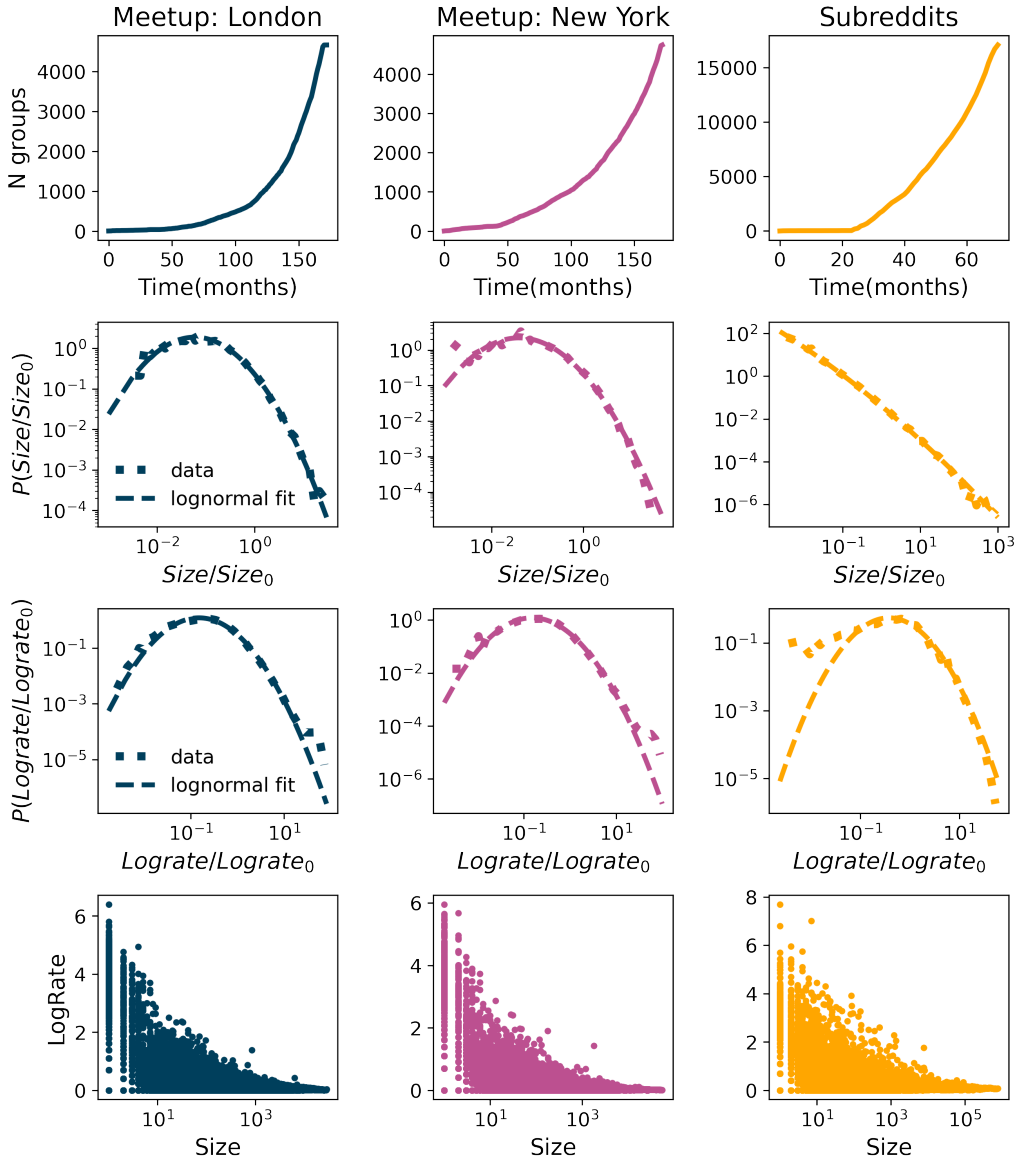

Figure 4.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011 that were active in 2017.

Even though Meetup and Reddit are different online platforms, we find some common properties of these systems; see Figure 4.1. 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 lognormal distribution:

$$P(S) = \frac{1}{S\sigma\sqrt{2\pi}} exp(-\frac{(\ln(S)-\mu)^2}{2\sigma^2}) \tag{4.1}$$

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 a similar lognormal 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 lognormal 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 lognormal distribution is a better fit for these data than the power-law. The Result section is given a detailed analysis that supports these findings.

The simplest model that generates the lognormal distribution is the multiplicative process [80]. 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 implies the lognormal distribution of the sizes, while the distribution of growth rates appears to be a normal distribution, [97], [98]. Figure 4.2 shows the distribution of the logrates that follow a lognormal distribution, contrary to the Gibrat law. Furthermore, logrates depend on the group size 4.2. For these reasons, the Gibrat law can not explain the growth of online social groups. Similar conclusions are shown in recent studies about cities or the growth of the internet [99, 100].

The growth of online social groups has universal behavior independent of the group's size. If we aggregate the groups created in the same year y, and each group size normalizes 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 4.2. The same characteristics are observed for the distribution of the normalized logrates 4.2. The growth is universal over time, and the group sizes distribution does not change from year to year.

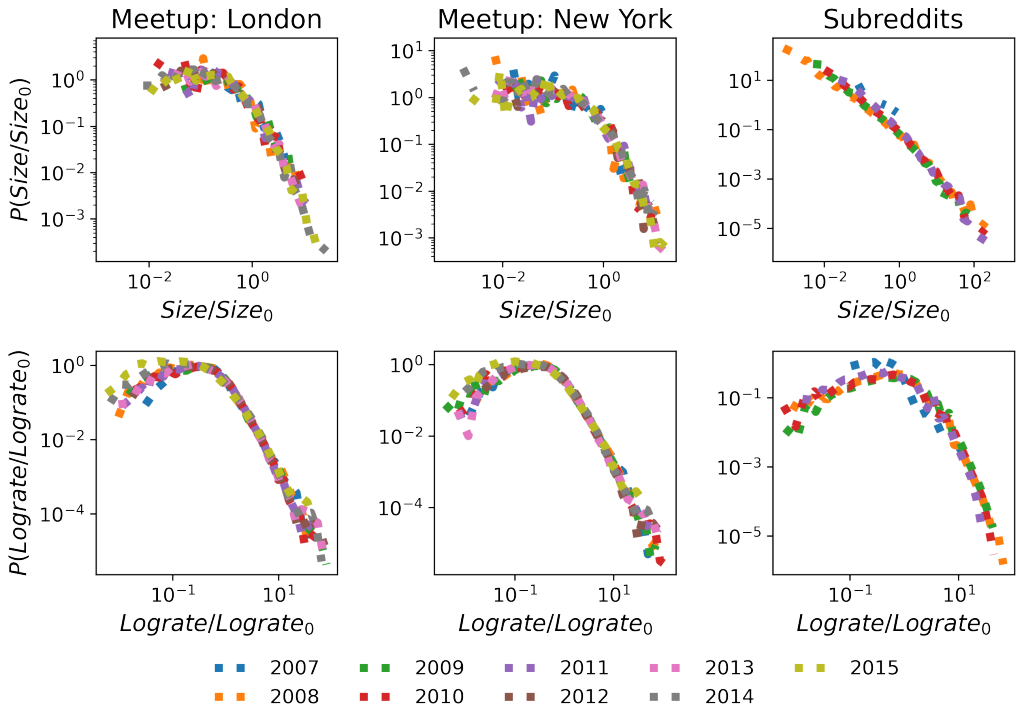

Figure 4.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

4.2 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 an arbitrary number of groups. In the previous section, we identified the universal patterns in the growth of social groups, but the growth can not be modelled with the 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 growth by varying linking rules. Regarding the user's group choice, it was shown that social connections play an important role [101, 102]. 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 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 [103, 104, 52].

The evolution of the social groups has been studied using the co-evolution model in the reference [102]. 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 lognormal, 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.

4.2.1 Groups growth model

The representation of the model is given in Figure 4.3. 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. New users $N_U(t)$ are added to the network at each step. 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 [101]. 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 4.3. When the user creates a new group or joins an existing one, the link $e(u, g)$ is made in the bipartite network B.

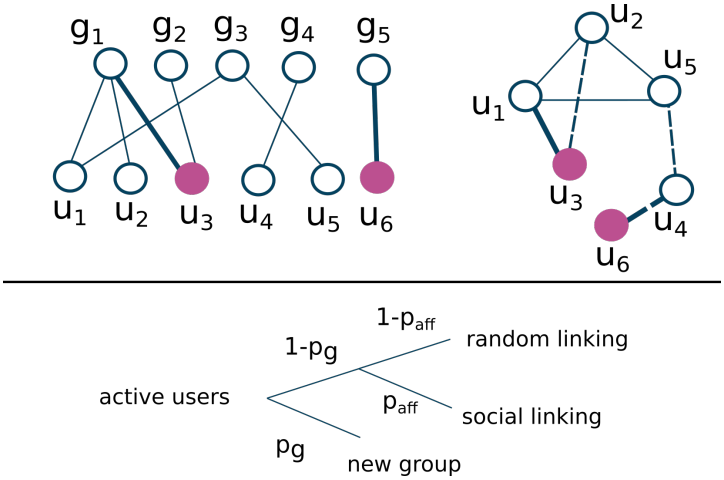

Figure 4.3: The top panel shows bipartite (member-group) and social (member-member) networks. Filled nodes are active members, while thick lines are new links in this time step. In the social network, dashed lines show that members are friends but do not share the same groups. The lower panel shows the model schema, where p_g is the probability that the user creates a new group, while p_{aff} is the probability that group choice depends on social connections. **Example:** member u_6 is a new member. First, it will make a random link with node u_4, with probability, p_g makes a new group g_5. With probability, p_a member u_3 is active, while others stay inactive for this time. Member u_3 will, with probability $1 - p_g$ choose to join one of the old groups, and with probability p_{aff} linking is chosen to be social. As its friend u_2 is a member of group g_1, member u_3 will also join group g_1. When member u_3 joins to group g_1, it 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.

$$P_{ug} = \sum_{u_1 \in \mathcal{N}_g} A_{uu_1} \tag{4.2}$$

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 [94, 105, 102] 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 a group g and make new connections in the 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.

4.2.2 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 the user is active to $p_a = 0.1$, so we can, in more detail, explore the influence of parameters p_g and p_{aff}. We plot the group size distribution after the 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 a large number of groups, over 10^4, while with $p_g = 0.5$, they are on the order of magnitude 10^5.

Figure 4.4 show the obtained group size distributions with power-law and lognormal fits. Users join randomly chosen groups for a lower parameter value $p_g = 0.1$ and $p_{aff} = 0$. Group size distributions are approximated with lognormal. When the affiliation parameter is larger, $p_{aff} = 0.5$, the lognormal 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 lognormal, but it is not explained with power-law either, right column on Figure 4.4.

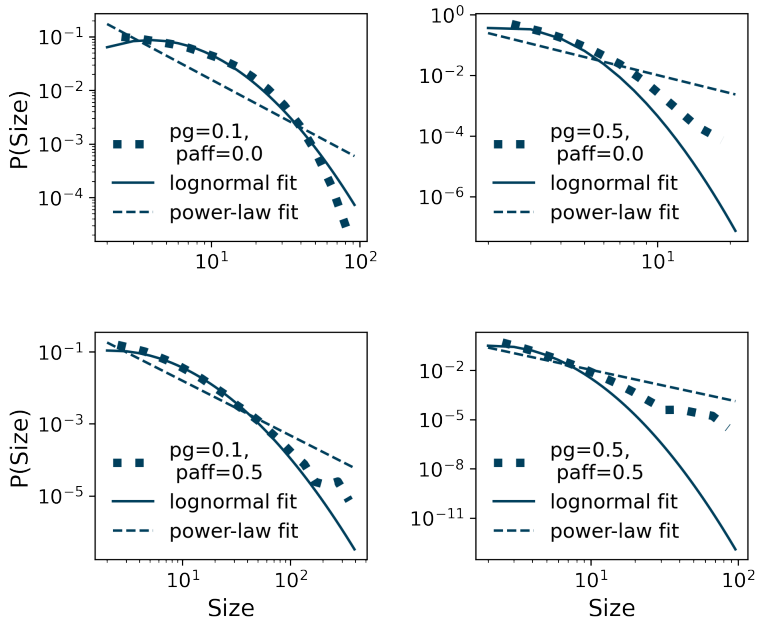

Figure 4.4: The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

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 [102], probability depends on the group size, as in the preferential attachment model. Instead of random linking, if we incorporate preferential linking, users with probability $1 - p_{aff}$ tend to choose larger groups, and group size distribution changes significantly. Similar to the co-evolution model, we find the power-law distribution. Figure 4.5 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 the affiliation parameter takes values 0, 0.5 and 0.8. If we consider random linking, a top panel on Figure 4.5, 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.

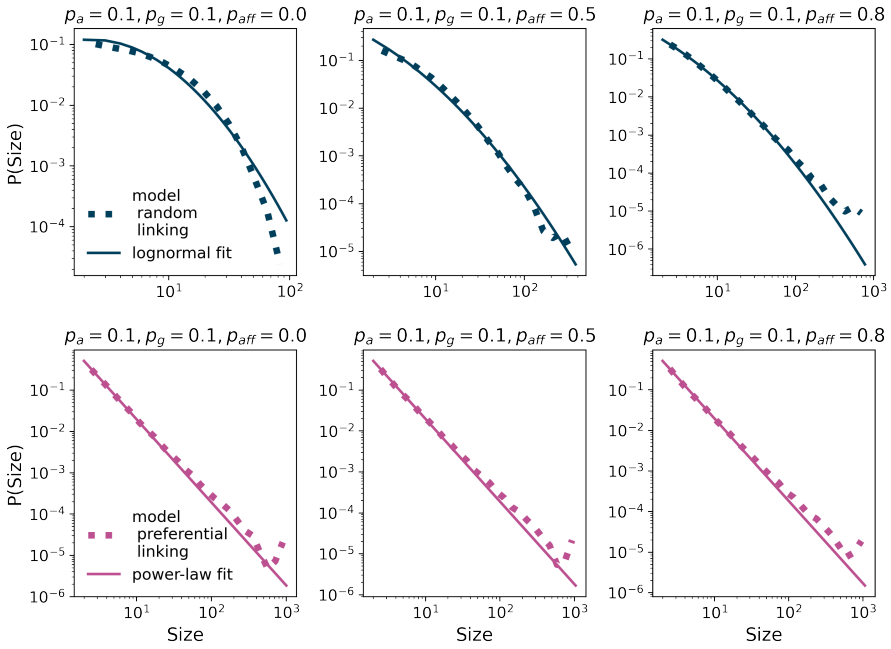

Figure 4.5: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff}. With probability, $1 - p_{aff}$, users choose a group randomly. The group sizes distribution (top row) is described with a lognormal distribution. The distribution has a larger width with a higher affiliation parameter, p_{aff}. The bottom row presents the case where with probability $1 - p_{aff}$, users prefer larger groups. For all values of parameter p_{aff}, we find the power-law group sizes distribution.

4.3 Results

The social systems do not grow at a constant rate. In Ref. [52], 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 4.6 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 initially, which accelerates as the system becomes more popular.

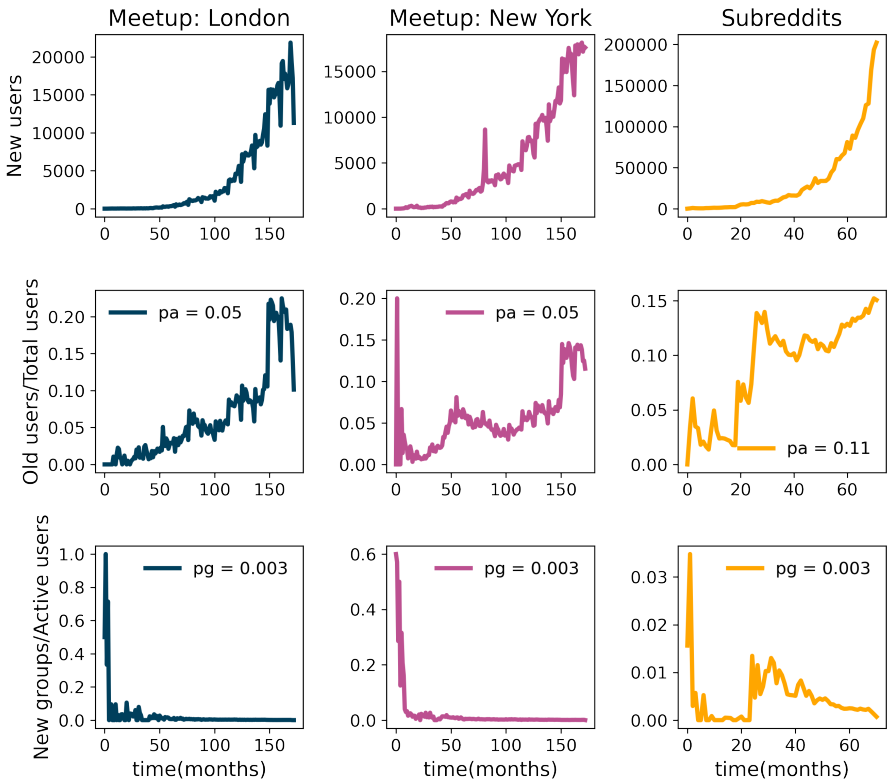

Figure 4.6: The time series of the number of new members (top panel), the ratio between old members and total members in the system (middle panel), and the ratio between new groups and active members(bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

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 active in month t and the total number of members in the system at time t. Figure 4.6 middle row shows the variation of parameter p_a during the considered time interval for each system. The values of this parameter fluctuate between 0 and 0.2 for London, and New York-based Meetup groups, while for Reddit, it ranges between 0 and 0.15. 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 the 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 4.6 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 Figure 4.6 that $p_g(t)$ has relatively high values at the beginning of the system's existence. In the beginning, these systems have a relatively small number of groups and often cannot meet the needs for the content of all

their members. As time passes, the number of groups and content offerings within the system grows, and members no longer have a high need to create new groups. Figure 4.6 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.

The affiliation parameter p_{aff} cannot estimate directly from the empirical data. For these reasons, we simulate the growth of social groups in 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 [106] between two distributions P and Q is defined as

$$JS(P,Q) = H\left(\frac{P+Q}{2}\right) - \frac{1}{2}\left(H(P)+H(Q)\right) \tag{4.3}$$

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 JS divergence value, the better the match between empirical and simulated group size distributions. Table 4.1 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 Reddit members, their social connections are more important regarding the 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 4.1: Jensen Shannon divergence between group sizes distributions from model (in the model, we vary affiliation parameter paff) and data.

Figure 4.7 compares 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 model well reproduces the tail of the distribution. The bottom row of Figure 4.7 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}.

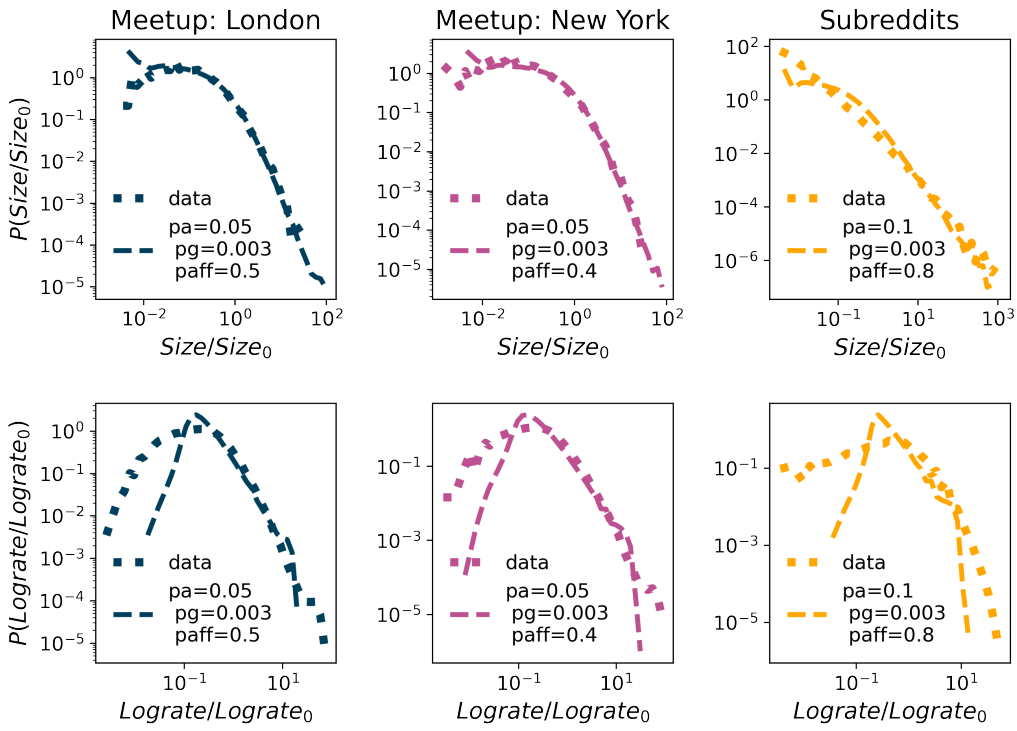

Figure 4.7: The comparison between empirical and simulation distribution for group sizes (top panel) and lograte (bottom panel).

4.3.1 Distributions fit

We compute the log-likelihood ratio R and p-value between different distributions and lognormal fit [85] to determine the best fit for the group size distributions. Distribution with a higher likelihood is a better fit. The log-likelihood ratio R has a positive or negative value, indicating which distribution represents a better fit. To choose between two distributions, we need to calculate the 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 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 4.2 summarizes the findings for empirical data on group size distributions from Meetup groups in London and New York, and Reddit. Using the maximum likelihood method, we obtain the parameters of the distributions [**?**]. The results indicate that lognormal distribution best fits all three systems. Figure 4.8 shows the distributions of empirical data and lognormal fit on data. For Meetup data, we present fit on stretched exponential distribution, which fits a large portion of data well. For subreddits, distribution is broad and potentially resembles power-law. Still, the lognormal distribution is a more suitable fit.

We use the same methods to estimate the fit for simulated group size distributions on Meetup groups in London, New York, and Subreddits. Table 4.3 shows the results of the log-likelihood ratio R and p-value between different distributions. We conclude that lognormal distribution is most suitable for simulated group size distributions. We confirm our observations by plotting lognormal and stretched exponential fit on data, Figure 4.9.

Table 4.2: 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

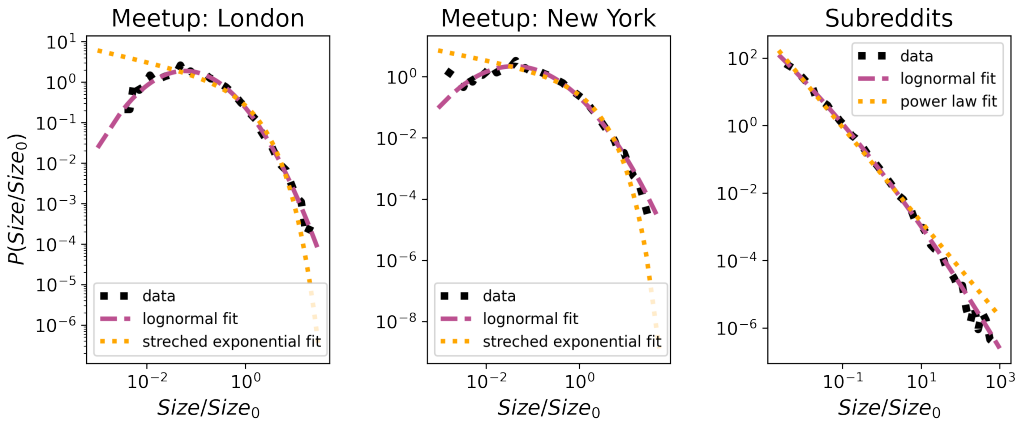

Figure 4.8: The comparison between lognormal and stretched exponential fit to London and NY data, and between lognormal and power law for Subreddits. The parameters for lognormal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits $\mu = -5.41$ and $\sigma = 3.07$.

Table 4.3: 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

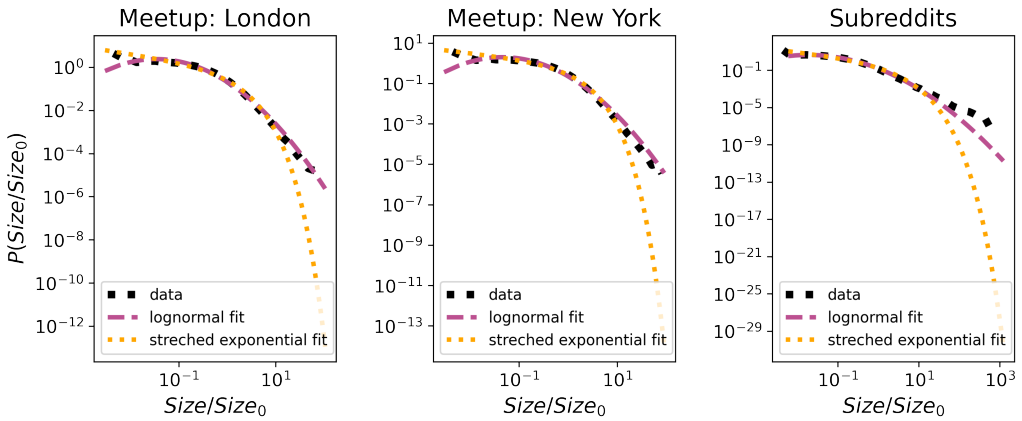

Figure 4.9: The comparison between lognormal and stretched exponential fit to simulated group size distributions. The parameters for lognormal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma = 1.53$.

4.3.2 Users partition in bipartite network - degree distribution

So far, the group growth model has focused on the degree distribution of groups and under what rules the universalities in the system reflected in the lognormal distribution of group sizes emerge. The model parameter p_a controls the users' activity level; otherwise, it shapes the degree distribution of users in the bipartite network. As this probability is constant and uniform among all users, we do not expect rich properties of users' degree distribution. The expected distribution is exponential for growing random graph [108], and the groups' growth model produces the same property. In Figure 4.10, blue dots show degree distributions of modelled Meetup and Reddit systems. This distribution is very well fitted with exponential form. Furthermore, in empirical data, these distributions are long-tailed, green dots in Figure 4.10, so the model can not reproduce the degree distribution of the users. In real systems, the probability that the user is active does not have to be uniform and constant. The previous work proposed that each user has a specific lifetime [109], but different linking rules could play an important role in shaping users' degree distribution. For example, p_a could be preferential toward high-degree users or even be time-dependent.

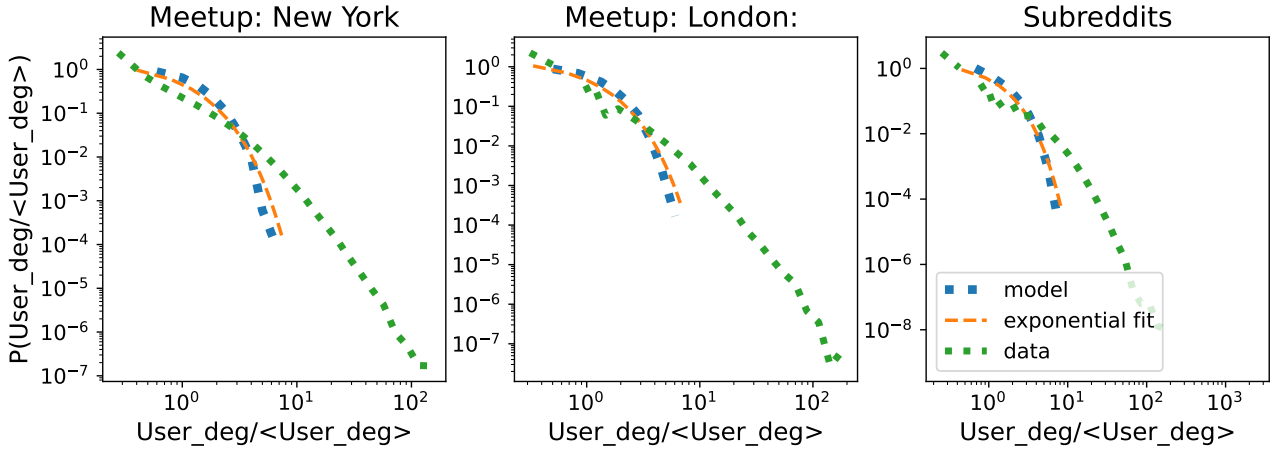

Figure 4.10: Users degree distribution

4.4 Conclusions

We apply complex network theory and statistical physics methods to describe the evolution of online social groups, Meetups in London and NewYork and Reddits. Instead of studying user interaction networks in a single group, which is a common approach, we are interested in quantifying how users interact with the system of multiple groups and determining which processes drive the growth of groups. Similar systems have been analyzed before. For example, it was found that the distribution of the cities or firms follows the lognormal and stays stable, showing universal behaviour. Contrary, the previous work on online social groups indicated that group size distributions of LiveJournal and Youtube follow power-law [102]. On the other hand, for Meetup and Reddit, we find the emergence of lognormal distribution of group sizes, and the distribution of Reddit is much broader. Furthermore, these systems grow exponentially in the number of groups and new users.

Meetup and Reddit may be platforms with different purposes, but on the lower level, both systems could be described with the same processes users perform: they can join existing groups or create new ones. Also, in these systems, new users constantly arrive. As we find the lognormal distribution in group sizes, our first attempt was to describe this system with the Gibrat model. It is a proportional growth size model, where group size distribution converges to the lognormal distribution while the log rates take the normal distribution. The second condition still needs to be met, so we need to use a more intricate method.

To explore the growth of these systems in more detail, we use a model where the social system is presented with evolving bipartite and social networks [102]. The bipartite network has partitions of users and groups, and a link exists if a user is a group member. The social network describes the social connections between members. At each time step, new users arrive in the system, following the time series of new users, and with probability, p_a old members decide to be also active. The active users can create a new group with probability p_g; otherwise, they will join existing groups. Their decision to select a group based on social connection is determined with probability p_{aff}; otherwise, the choice is random.

We estimate model parameters p_a, p_g and p_{aff} from empirical data. We see that model approximates well the empirical distributions. For Meetup groups in London and New York, the p_{aff} parameter is smaller, while for Reddit, p_{aff} is higher, resulting in broader group size distribution. It also means that for Reddit members, social connections are more important for the choice of groups.

With results in this chapter, we contribute to the knowledge of the growth and segmentation of the socio-economic systems. Our work was motivated by the Co-evolution model [102]. The authors explore the social groups in which group size distribution scales as power-law. We identified different universality class, the system where group size distribution follows log normal. Further, we marked off a set of linking rules which led to lognormal group size distribution and compared these two cases. By this, we expanded the classes of social systems that can be modelled.

Chapter 5

The role of trust in knowledge-based communities

The **Stack Exchange** (SE) 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 Stack Exchange network, which 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 can 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. However, SE criteria for graduation have not been applied consistently on every SE site, as many sites graduated without reaching all required thresholds. As those measures only quantify the overall number of questions, answers or highly active users, we want to understand how SE community structure evolves and identify factors that influence sustainability. The need to share knowledge with others motivates users to use Q-A platforms. Still, the fact that they interact with each other reveals their sense of belonging to the community and the presence of trust among users. Our proxy for measuring trust in the community is the Dynamic Interaction Based Reputation Model.

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 the similar type of users.

[1] More information about StackOverflow is available at https://stackoverflow.co/, and a broad introduction to the SE 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.

5.1 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 multiple interactions between users or the direction of the interaction.

The first approach is to aggregate all interactions in the first 180 days and study the properties of the static network. Many local and global network measures are dependent [40], and it was shown that degree distribution, degree-degree correlations and clustering coefficient are sufficient for the description of the properties of complex networks [110].

We calculate the **degree distribution**, Figure 5.1, and compare the distributions of active and closed communities of the same topic. Degree distributions between active and closed communities follow similar lines.

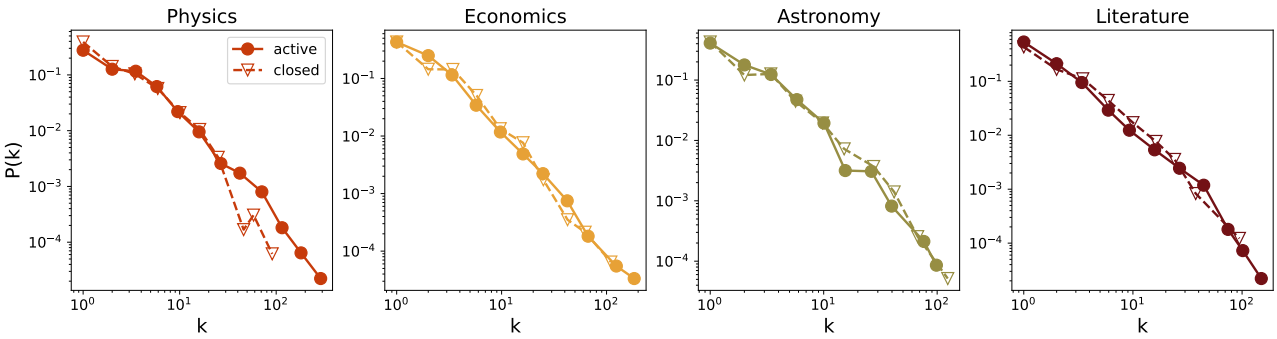

Figure 5.1: Degree distribution.

If we take a look into **neighbor degree** depending on the node degree $k_{nn}(k)$, Figure 5.2, 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 neighbours with a larger degree than is the case in closed communities. The results are consistent for Physics, Economics and Literature. For Astronomy, we find different behaviour, where the $k_{nn}(k)$ distributions of closed communities are on top of the distributions of the active ones.

The **clustering coefficient** of a node quantifies the average connectivity between its neighbours and cohesion of its neighbourhood [40]. It is a probability that two neighbours of a node are also neighbours and is calculated using the following formula:

$$c_i = \frac{e_i}{\frac{1}{2}k_i(k_1 - 1)} \,. \tag{5.1}$$

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. A study on dynamics of social group growth shows that links between one's friends that are members of a social group increase the probability that that individual will join the social group [105]. Furthermore, successful social diffusion

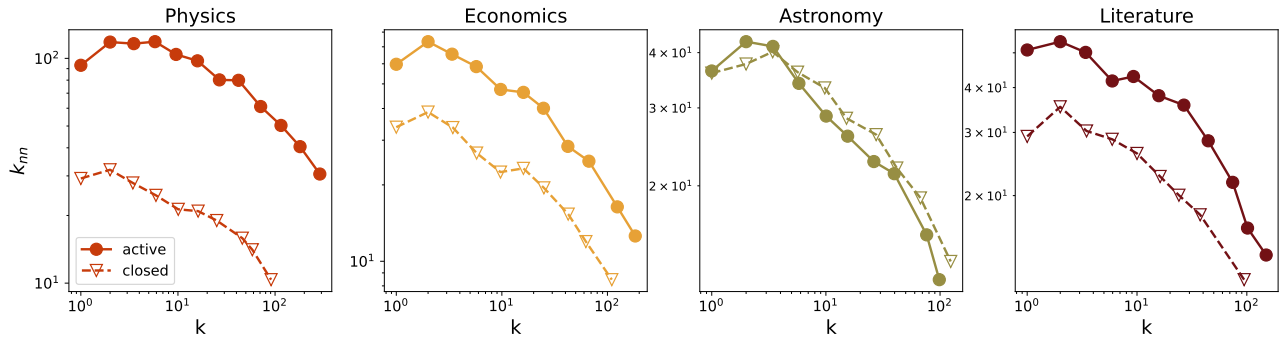

Figure 5.2: Neighbour degree.

typically occurs in networks with a high value of clustering coefficient [111]. 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 in Figure 5.3. As expected, we find that active communities are more clustered.

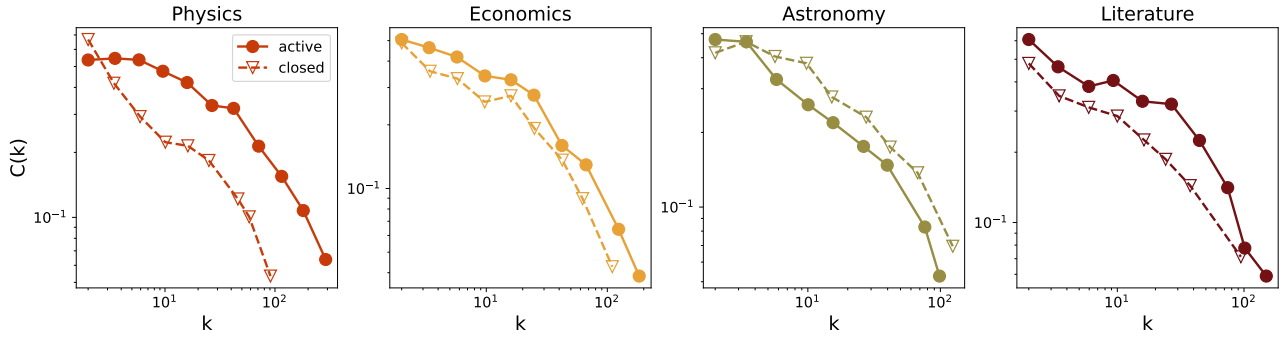

Figure 5.3: Clustering coefficient.

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 the time window of the length τ. We fix the time window to $\tau = 30$ days and slide it by $t = 1$ day through time. A 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 of consecutive networks overlap significantly.

Here we investigate how the SE community's clustering coefficient changes with time by calculating its value for all network snapshots. We compare the behaviour of clustering for active and closed communities on the same topic to better understand how the cohesion of these communities is changing over time. Figure 5.4 shows the evolution of the mean clustering coefficient for all eight communities. All communities still alive are clustered, with the value of the mean clustering coefficient higher than 0.1. Physics, the only launched community, has a clustering coefficient value above 0.2 for the first 180 days.

During the larger part of the observed period, an active community's clustering coefficient is higher than its closed pair's clustering coefficient. Let's compare active communities with their closed counterpart. The closed communities have a higher value of the mean clustering coefficient in the early phase, while later communities that are still active have higher clustering coefficient values. These

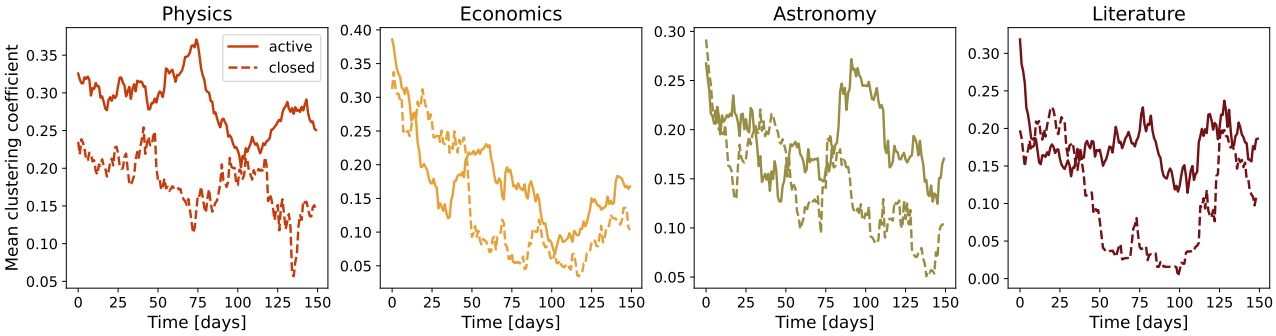

Figure 5.4: Mean clustering coefficient.

results suggest that all communities have relatively high local cohesiveness and that lower clustering coefficient values may indicate its decline in the later phase of community life.

5.2 Core-periphery structure

Previous research on Stack Exchange communities has attempted to explain how different types of users interact. In Question-Answer communities are expected to be popular and casual users [112, 113]. Popular users generate the majority of interactions in the system; they are experts in the community and take care of answering questions and engaging the discussions through comments. As popular users, they considered the 10% of the most active users and showed that popular users are highly connected with themselves and casual users.

We tested this theory on all eight communities. We focused on 30 days of sub-networks and showed how the Number of links per node among popular users and between popular and casual users evolves, Figure 5.5. We also compare active and closed communities of the same topic, so links per node in active sites are more significant than in closed communities.

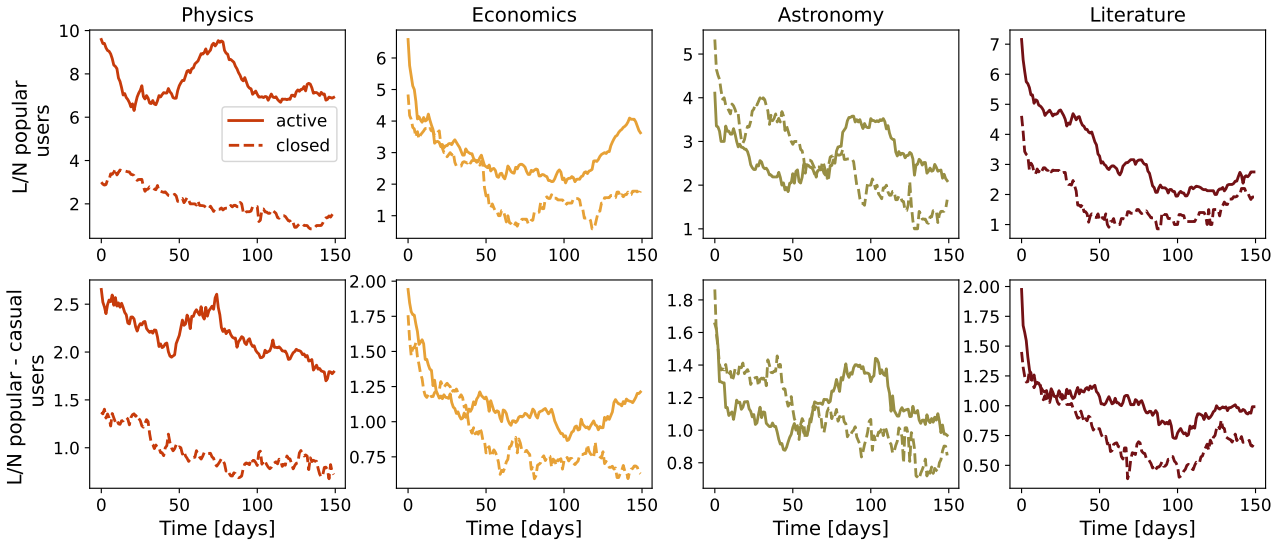

Figure 5.5: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users).

Although we find the difference between active and closed communities, the split according to 10%

most active users does not guarantee that all popular users will be considered. Furthermore, the smaller group of frequently active users is similar to the core users in the core-periphery structure. This is why we will detect the core of each 30-day network. By this, separation is based on the network structure and is more consistent, as using the algorithmic approach, we optimize the connectivity inside the core, periphery and among them. The core-periphery structure has a core that is a densely connected group of nodes, while the periphery has a low density [64, 55].

We use the Stochastic Block Model (SBM) to infer the core-periphery structure of each 30 days network snapshot and analyses how the core structure evolves. The SBM algorithm is adapted for inferring the core-periphery structure, [55]. For each 30 days network, we run the sample of 50 iterations and choose the model parameters according to the 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 the minimum description length does not fluctuate much. Also, different similarity measures between inferred core configurations take values higher than 0.9, indicating that the core structure is stable.

The Number of users in the core of active communities is higher than in closed communities, the top panel on Figure 5.6. On the other hand, we do not find a big difference between the fraction of core users in the closed and active communities. Furthermore, the fraction of users in core differs from the 10%, and it is constantly changing, bottom panel 5.6.

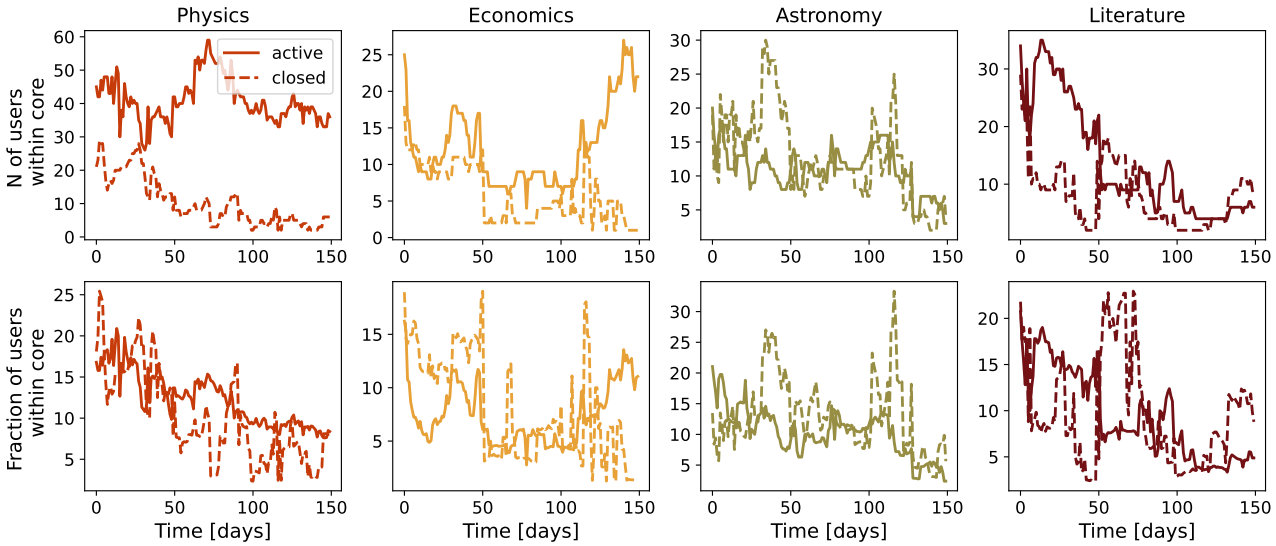

Figure 5.6: The size of the core (top) and a fraction of users in the core (bottom). Solid lines - active sites; dashed lines - closed sites.

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 the Jaccard index indicate the more similar cores. The highest values are found around diagonal elements where we compare networks closer in time, see Figure 5.7. The core membership changes over time, and the change is more frequent in 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 5.8. The Jaccard index decreases with the relative time difference between networks faster in closed communities. The relatively high overlap between distant networks confirms that active networks have a more stable core.

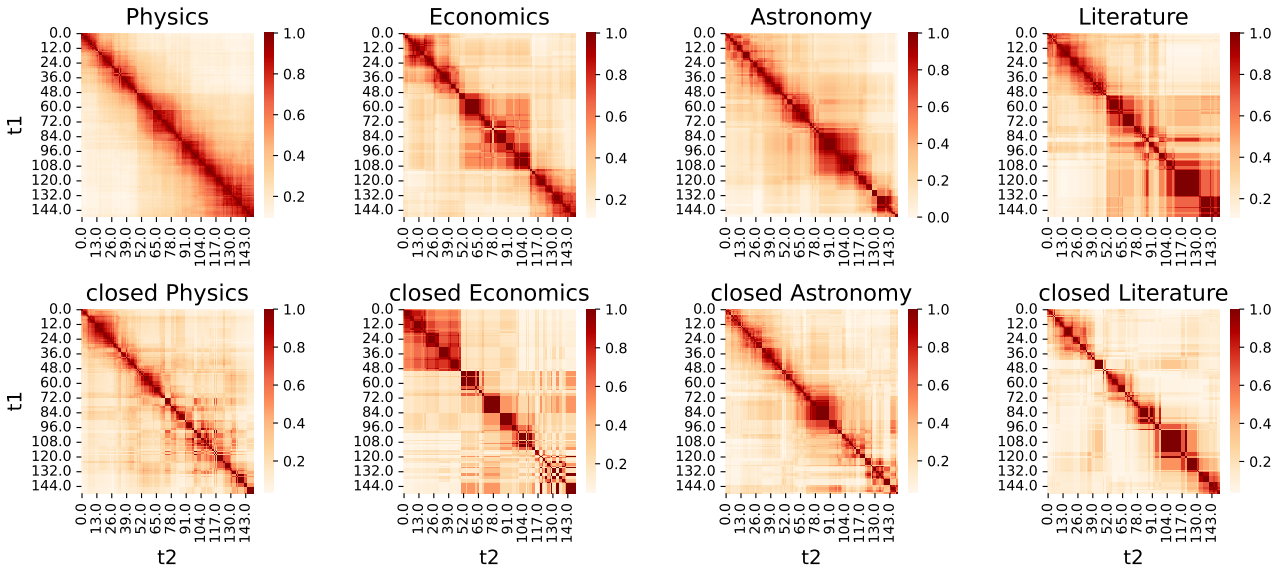

Figure 5.7: Jaccard index between core users in sub-networks at time points $t1$ and $t2$.

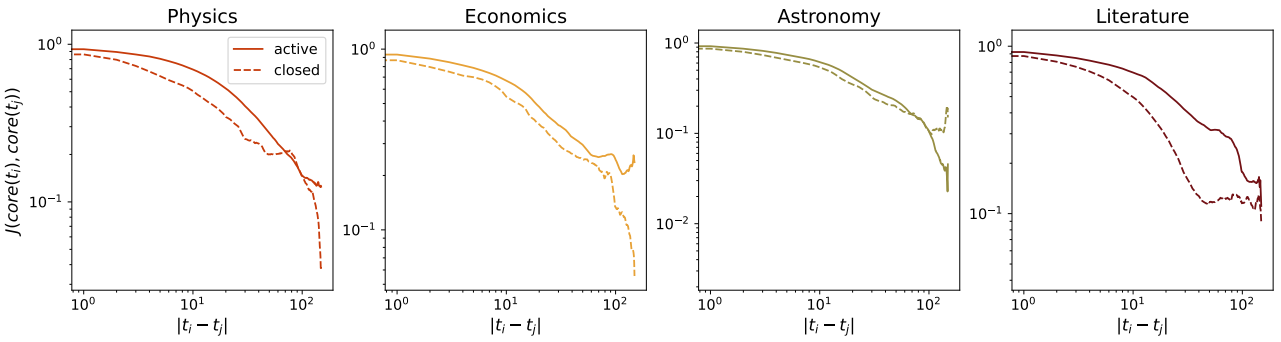

Figure 5.8: Jaccard index between core users in 30 days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$.

Finally, we examine how the users' connectivity in and between the core and periphery evolves. In the Figure 5.9, we show the L/N in the core, which is proportional to the average degree of the network $2L/N$, . The Physics community has more than twice the connectivity than closed Theoretical Physics. For Literature, we also find higher connectivity. Still, at the end of the observation period, the connectivity in the active site drops and becomes similar to that in the closed one. The difference between active and closed sites is unclear for Economics and Astronomy. At the beginning of the period, connectivity is similar for the sites on the economic topic. After 50 days of community life, connectivity in active communities is starting to rise, while in the case of closed economics, it is dropping. Astronomy connectivity is higher in closed communities in the first 50 days. After this period, we find a sudden rise in the connectivity of active astronomy, but again it drops and becomes comparable to the connectivity values in the closed site. Similar conclusions can be drawn for the connectivity between the core and periphery. The largest difference between active and closed sites is observed in Physics.

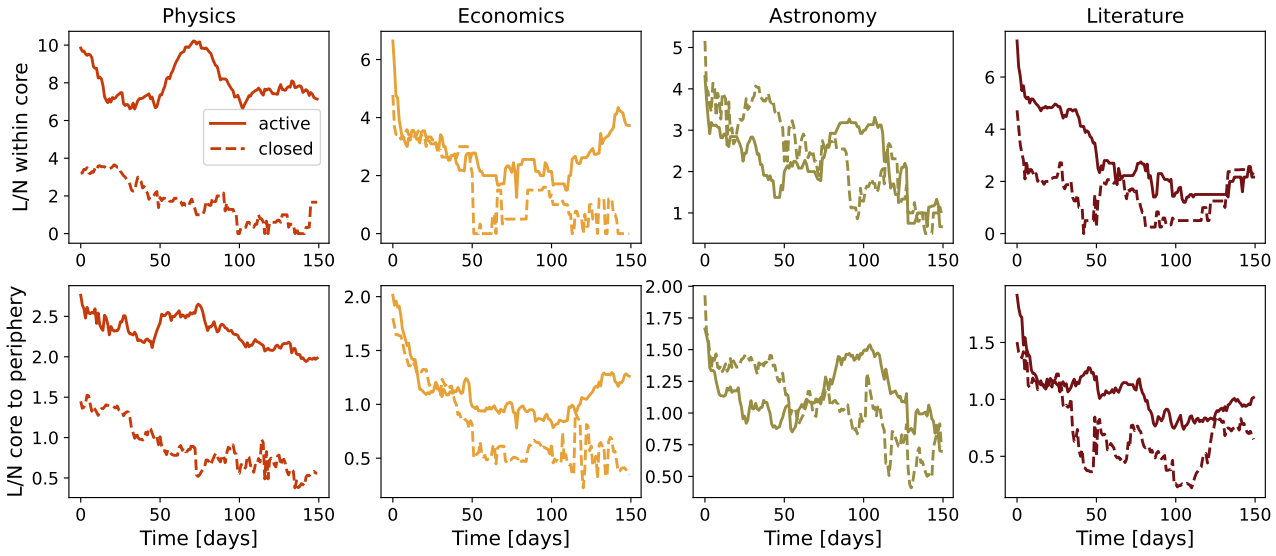

Figure 5.9: Links per node in core and links per node between core and periphery.

5.3 Dynamical Reputation on Stack Exchange communities

We further explore the difference between active and closed communities through the dynamic reputation model. With this model, we calculate each user's reputation in the community, and reputation is directly connected with the collective trust in the network.

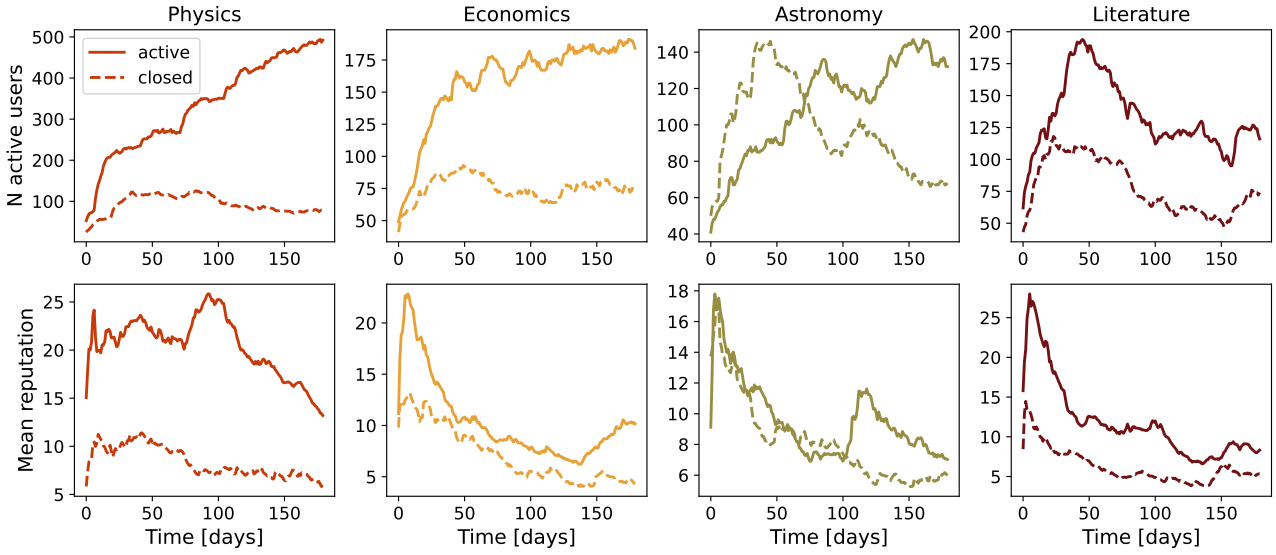

Figure 5.10: Dynamic Reputation on the four pairs of Stack Exchange websites: Astronomy, Literature, Economics, Physics and Theoretical Physics.

Dynamical reputation model, introduced in section 2.6, has three parameters. We explored different parameter combinations to find the set of parameters the most suitable for a given system of Stack Exchange communities. First, the basic reputation is set to $I_{bn} = 1$. The cumulative factor is $\alpha = 2$, as we want to emphasize the frequent interactions. The parameter β controls the reputation decay due to user inactivity. After the last activity, the user has a positive reputation for some period

and is still impacting the other users. We optimized the Number of users with a reputation larger than 1 according to the number of users in the 30 days network and concluded that parameter $\beta = 0.96$. A more detailed discussion about the choice of parameters is in the appendix.

With selected model parameters, we calculated the reputation of each user. If a user has a reputation larger than 1, it is considered active, but when the reputation drops below this threshold means that the user has not been active long enough; it does not make a valuable contribution to the community. The Number of active users and their mean reputation for different SE sites is shown in Figure 5.10.

From the properties of networks, we found that active communities are more cohesive and have a more stable core. Furthermore, we focus our analysis on the dynamic reputation of the core users. Figure 5.11 shows the evolution of mean user reputation within the core. Active communities have a larger reputation than their closed counterpart. As it is previously suggested, the largest difference is found in the Physics community. For other communities, the difference is not so striking; on average, the core of active communities has a larger reputation than the core of closed communities.

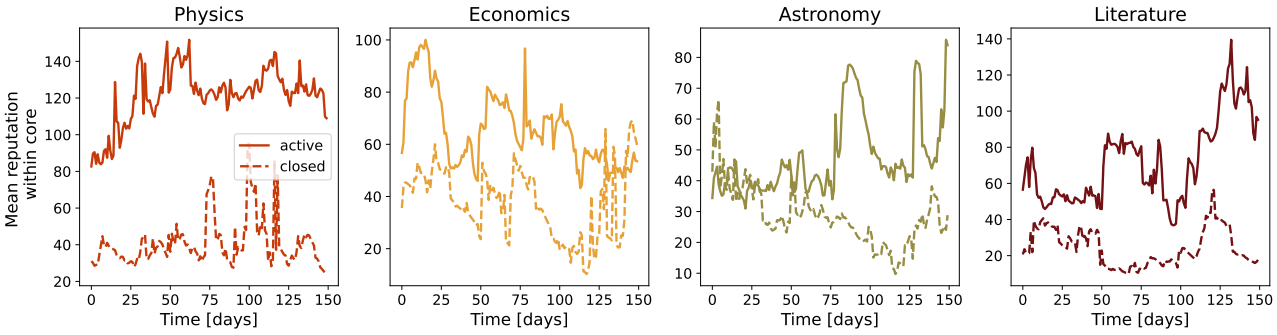

Figure 5.11: Dynamical reputation within the core.

In the network's core are active users, and we expect a higher dynamic reputation than the total reputation of users belonging to the periphery. The ratio between core and periphery in Physics is always higher than in Theoretical Physics, and similar conclusions are observed in the Literature. In the early days of Economics, we find a different pattern; the core-periphery reputation ratio is larger for closed Economics, but later it changes in favour of active Economics. Astronomy shows different behaviour where the closed community where dominant; closed astronomy had a larger core-periphery reputation ratio.

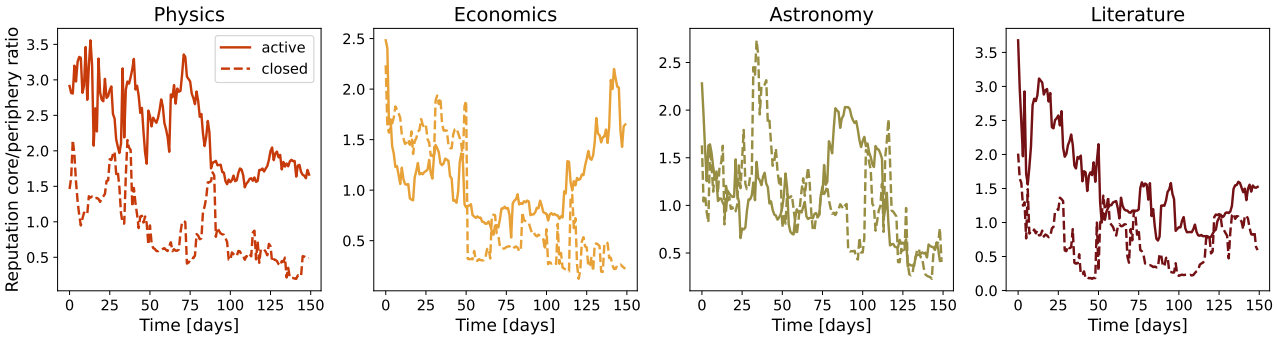

Figure 5.12: Ratio between the total reputation within network core and periphery. Solid lines active communities, dashed lines closed communities.

The distribution of the dynamic reputation of SE communities is skewed. We calculated the Gini coefficient to better express the difference between distribution reputations. This measure quantifies the inequality among users' reputations. The Gini coefficient is calculated based on reputation values for each day; see Figure 5.13. The Gini coefficient is larger than 0.5 in the first 180 days. Also, the active communities showed more reputation inequality, and dynamical reputation has a larger variation.

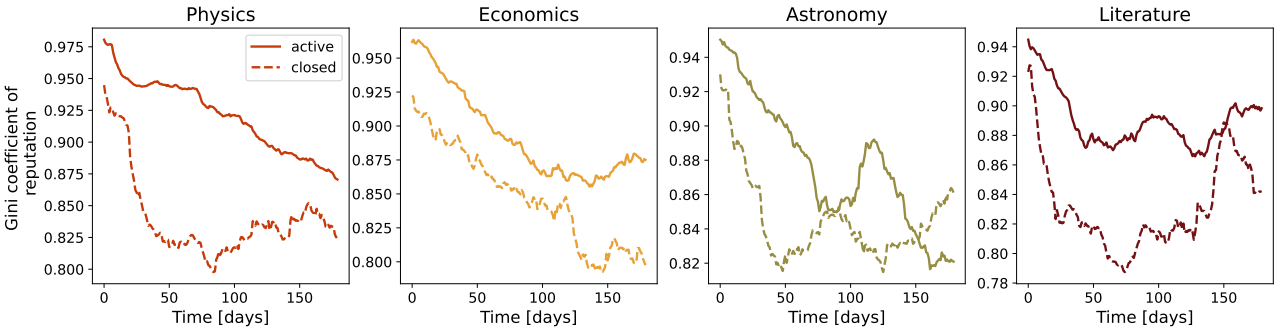

Figure 5.13: Gini index of dynamic reputation within the population.

Further, we investigate how the properties of user interaction networks correlate with the user's reputation. For example, we can measure the assortativity coefficient among connected users in the network. For each 30 days user interaction network, we calculate the reputation assortativity, using the reputation value observed on the last day of the time window in which the network is constructed. With this measure, we quantify whether users tend to connect with users with similar reputations or not. Figure 5.14 shows results where we compare each SE community's active and closed sites. Assortativity has small values in all communities' reputations, not larger than $|0.3|$. In active communities, this is a mostly negative measure showing expected user behaviour: popular users, who often have a high dynamical reputation, interact with users with low dynamical reputations. Astronomy is an outlier again; during the first 100 days active community had a positive reputation for assortativity, and after this period, it started behaving similarly to other active communities.

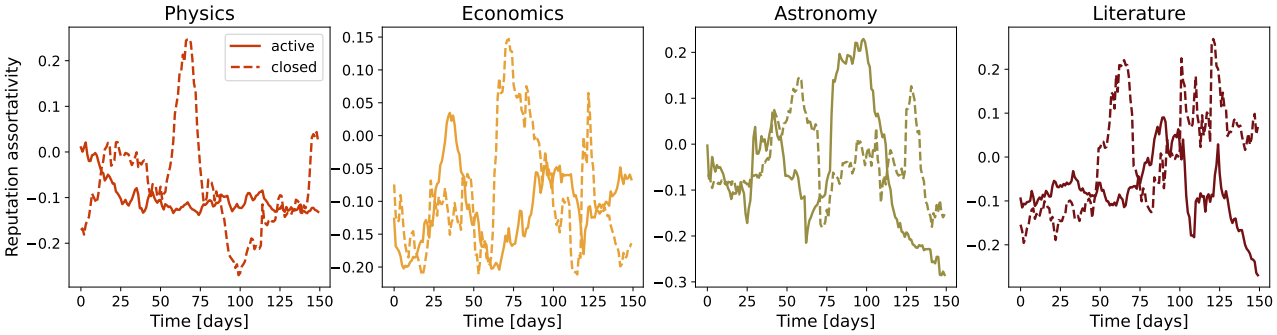

Figure 5.14: Dynamic Reputation assortativity in the network of interactions (questions, answers, comments, unweighted, undirected network). Solid lines - active sites; dashed lines - closed sites.

Finally, we are interested in how dynamical reputation correlates with network measures. We compare the node's centrality in the 30-day network and the node's reputation on the last day of the 30-day sliding window. The correlation coefficient between dynamic reputation and node degree is very high; see the top panel on 5.15. The bottom panel shows correlations between dynamic reputation and betweenness centrality in the network, which are also high. We find that correlations are mostly

higher in active communities; only for astronomy do they take similar values during the observed period.

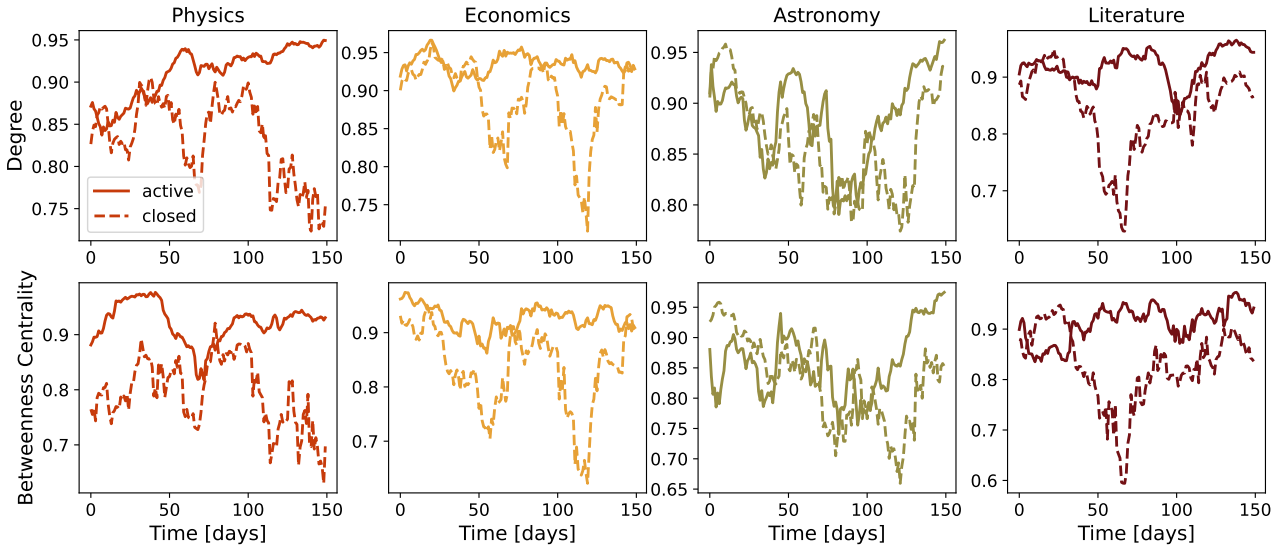

Figure 5.15: Coefficient of correlation between users' Dynamic Reputation and users' network degree (top) and users' betweenness centrality (bottom). Solid lines - active sites; dashed lines - closed sites.

5.4 Conclusions

The Stack Exchange sites bring together users interested in knowledge sharing. They create different topic communities where each member can post topic-related questions and get the correct answer from other users. The SE developed, in one sense, the trust among users, as many people see the SE as a valuable source of knowledge and seek their answers directly in these communities. Not all SE sites were launched, and some were closed because they did not fulfil the Stack Exchange criteria of the successful community. These criteria rely on basic measures such as the number of active users, posted questions, and answers, so we were interested in investigating the structure and dynamics of SE communities to understand how trustworthy and self-sustainable community emerges.

This chapter presented results on four pairs of SE communities: Astronomy, Literature, Economics and Physics. The first time each of them failed to create a sustainable network, but later the same topic was proposed communities are still active. While this sample may be small, we wanted to focus only on communities on the same topic, so our comparison between closed and active communities is not topic related. Also, we chose two communities from STEM and two from humanities which allowed us to remove field-related biases.

We studied how network properties evolve during the first 180 days. To closely examine the structure, we constructed the sub-networks within a 30days window. Sliding window by day, we continuously measure the structure of the network. The clustering coefficient is higher in active communities. The previous study suggested two groups of users in Q-A communities, popular and casual users [113]. This observation motivated us to closely analyze the network segmentation in the core-periphery structure. Based on Bayesian Stochastic modelling, we identify each 30-day network core user. Furthermore, using the DIBRM model[56], we quantify each user's reputation. This reputation is our proxy of trust, and its dynamics reflect some of the essential properties of trust. When a user is frequently active, the reputation increases; when inactivity declines, the user becomes less important.

Used methods have several parameters which need to be tuned according to specific systems properties. First of all, we showed that the choice of the sliding window does not influence our conclusions, as observed system properties follow similar patterns for different values of sliding windows. Tuning the DIBRM parameters was more changeling. Our primary assumption was that the number of users with a positive reputation should resemble the number in the 30-day window.

Our results suggest that core members are important for the sustainability of the community. The core members have a high reputation and contribute to the community's survival. The core is more connected in active communities, and larger connectivity is found between the core and periphery in active communities. The most noticeable difference between closed and active communities is in Physics. Physics is the only community that graduated after 90 days, while other active communities stayed in the beta phase for a couple of years; recently, their status changed to beta. On the other hand, closed Astronomy showed larger network properties than an active one, but as time progressed, this changed in favour of the active community. The larger mean reputation and its dynamics among core users in active networks are important indicators of a thriving community.

Chapter 6

Conclusions

In this thesis, we studied the complex network models to understand the evolution of online social systems. The complex systems change over time, even though we often find the system's collective behaviour that stays universal. The specific interactions among elements could lead to different kinds of organizational patterns. The research from this thesis tries to understand factors that contribute to the growth of the systems, structure properties and their sustainability. The underlying methodology is introduced in chapter 2. The first part explained the most important properties of network structure and the growing network models. The second part explains the statistical methods useful for the empirical analysis of the properties of the complex system.

In chapter 3, we discussed how nonlinear growth signal shapes the structure of the complex network. The previous models combined linking rules with constant growth, but we added one more parameter, the fluctuating growth signals, in this research. The most considerable influence is found on scale-free networks. Many interaction networks from social, technological or biological systems have scale-free structures; they are correlated and clustered. These results suggested that it is important to study growing signals' properties. Signals from natural systems show trends and cycles and are characterized by long-range correlations. The structure of the generated complex networks depends on the signal properties. It is necessary to quantify these properties as they affect the network's topology differently. For example, the most significant difference between networks generated with fluctuating and constant signals is found for signals with multi-fractal properties. This difference is more negligible for monofractal signals or uncorrelated white noise. Fluctuating signals promote the creation of hubs in the network and shorten the paths between nodes.

Chapter 4 presented the results of the universal characteristics of the growth of online social groups—the growth of the system influence the structure of the interaction network. The distribution of the sizes of the complex systems usually follows some universal curve. In many cases, it is lognormal or power-law. The distribution of the dimensions of the city sizes could be explained with Zipf law [114]. The number of citations scales as lognormal distribution [7]. In this thesis, we empirically analyzed the growth of online social systems. They consist of groups whose growth is universal. The empirical analyses of Meetup groups and Reddits showed their group size distribution follows universal lognormal distribution, stable over time. This research aimed to examine the structure and dynamics of the interaction network. We proposed the bipartite group model to gain a deeper understanding of the factors that affect the growth of social groups in a complex system. The growth in this model is driven by fluctuating signals, similar to the paper presented in chapter 2: we use a time series of new members from Meetup and Reddit. The number of groups also grows as each user can create

a new one; otherwise, the user joins the old group, and different linking rules determine his decision. The lognormal distribution of the group sizes emerges when with probability p_{aff}, users prefer groups whose friends are already members, while with probability $1-p_{aff}$, their choice is random. The width of the lognormal distribution depends on the parameter p_{aff}. The systems influenced more by social connections have larger p_{aff}, and the broader group sizes have lognormal distribution.

In chapter 5, we focused on the Question-Answers platform Stack Exchange. Each site goes through several phases before being successful and launched. During that period, the site may be closed without a strong community. We selected several topics in which sites for the first time were closed, but in the second attempt, they survived and are still active. We provide a detailed analysis of active and closed Stack Exchange sites, compare their properties and identify what is crucial for the community's survival. We map user interactions observed in 30 days onto complex networks. Further, we slide the window by one day and follow the evolution of the network.

According to the clustering properties of these networks, sustainable communities have a higher value of local cohesiveness. We use the Bayesian stochastic block modelling approach [55] to determine the core-periphery structure of these networks. We find that sustainable communities develop stable, better-connected cores. To analyze the evolution of collective trust in SE communities, we modify the Dynamic InteractionBased Reputation Model [56] (DIBR) model. We use the DIBR model to measure the user's reputation based on the frequency of their activity and its evolution during the first 180 days. The trust between core members of active communities develops early and is higher than in closed communities during the first 180 days. The early emergence of a stable, trustworthy core may be a crucial factor in determining a knowledge-sharing community's sustainability.

The question raised by this study is how trust emerges among users in questions answers communities where the users tend to share knowledge, and their communication is neutral or positive. Some communities started promoting hate speech on different online platforms, resulting in the banning. But, banned users remained in the online world; they moved their communities to alternative platforms without strict policies, such as Voat. Later, Voat users also formed no-hate speech topics, and there is an open question does the emergence of trust differ among different communities? On the other hand, exploring higher-order representations of online communities would be interesting. Threads, where more people reply to one post, could be studied using simplicial complexes to reveal complex network structure patterns. Furthermore, the research that employs agent-based modelling allows us to connect closer the actions of single users with the emergence of collective phenomena and the rise and fall of trust in the system.

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 the first 180 days of data. Given that the first few months can be crucial for the further development of the community [115], we are interested in the early evolution of Stack Exchange sites.

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

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

Even though posts can be voted on 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%

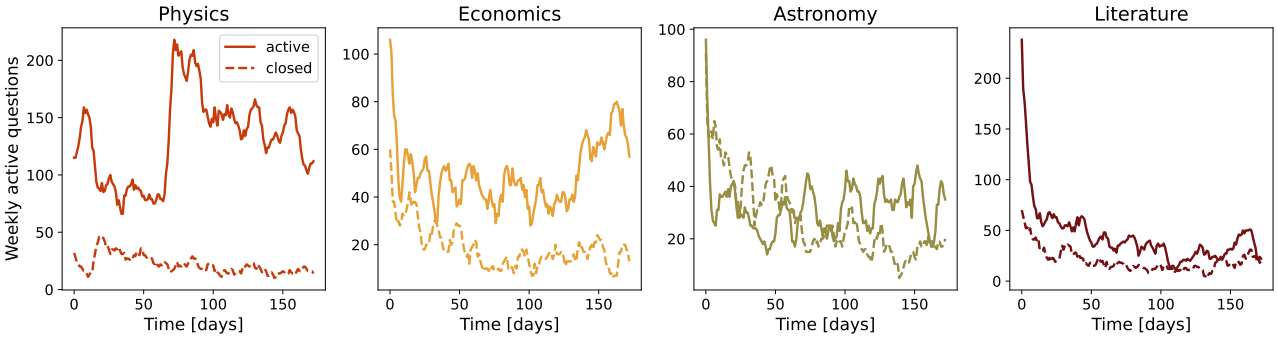

Figure A.1: Number of active questions within seven days sliding windows. Solid lines - active sites; dashed lines - closed sites.

A.1 Comparison between active and closed SE communities

Table A.2 compares the first 180 days between closed and active communities. Regarding basic statistics, active communities had a larger number of users, questions, answers and comments. Another simple indicator if the community will graduate or decline can be time series of active questions for seven days in Figure A.1. The question is active if it had at least one activity, posted answer, or comment during the previous seven days. We find that live communities have more active questions after the first three months. Still, this difference is smaller for literature and astronomy. For astronomy, we observe that closed communities had more 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 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 ten questions per day, 2.5 answers per question and more than 90% of answered questions. For less than 80% of answered questions, five 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 Table A.3. After 180 days, only live physics is a healthy site while other live communities are at least in two criteria labelled as *okay*. Closed sites mostly *need some work*; the

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

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

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

Appendix B

Selection of Dynamical Reputation Model parameters

The Dynamical Reputation Model(DIBRM) has several tuning parameters. In previous studies, the model [56, 116] was used to approximate real reputation on Stack Exchange sites [116], so model parameters were $t_a = 2, \beta = 1, \alpha = 1.4$, while the basic reputation value I_{bn} was +2 or +4. As $\beta = 1$, the forgetting factor is not considered. Our goal was to describe how reputation influences the sustainability of the community. Further, we wanted to resemble the concept of trust. Our tuning procedure differs from previous studies on Stack Exchange sites, and we ended up with different model parameters.

For **basic reputation contribution**, we selected $I_{bn} = 1$. With these values, each interaction has an initial contribution $+1$.

For **characteristic time** t_a we choose $t_a = 1$. The median/average time between subsequent interactions is $1day$. If the time window between two interactions is less than $1day$, their reputation will rise; otherwise, the reputation decays.

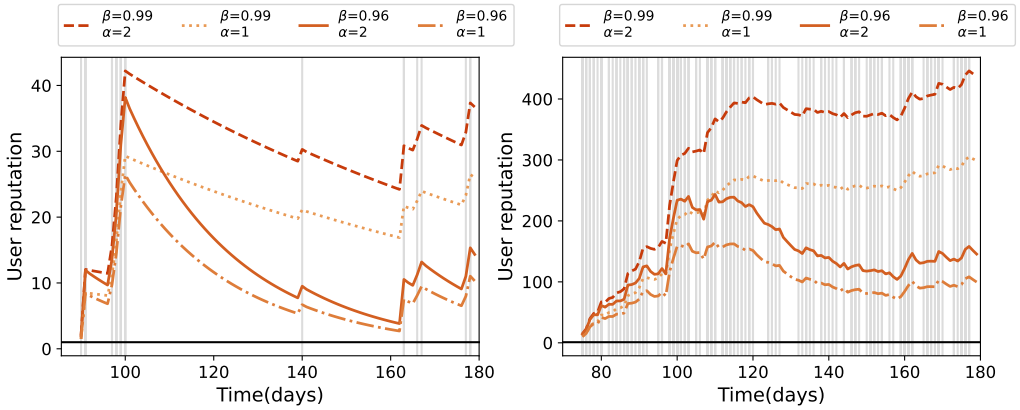

Figure B.1: Single users reputations

The parameter α represents the **cumulative factor**. The burst in activity and recent interactions lead to higher reputation values with larger parameter α. Figure B.1 represents the reputations of two

selected users from SE. The first is sporadically active, while the second makes frequent interactions. We calculate the reputation of these two users for different parameters (α, β). We selected $\alpha = 2$.

The reputation decay determines the **forgetting factor** β. We set the parameter on $\beta = 0.96$. The reputation should reflect the properties of the trust. This means we do not expect β to be high, as inactive users keep larger reputation values. In Figure B.1 for $\beta = 0.99$, even for the little active user, reputation stays higher during the observed period. With lower β, it may drop to the reputation threshold and indicate that the user stopped to be active.

We compared the number of users with an estimated reputation higher than 1 for different parameters β. We concluded that β close to 0.96 approximates the number of users with recorded interactions in a given 30-day sliding window. For each pair of communities, we calculated the number of users with at least one interaction in every 30-day sliding window. Then we estimated several times in series expressing the number of users with a 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 in 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.

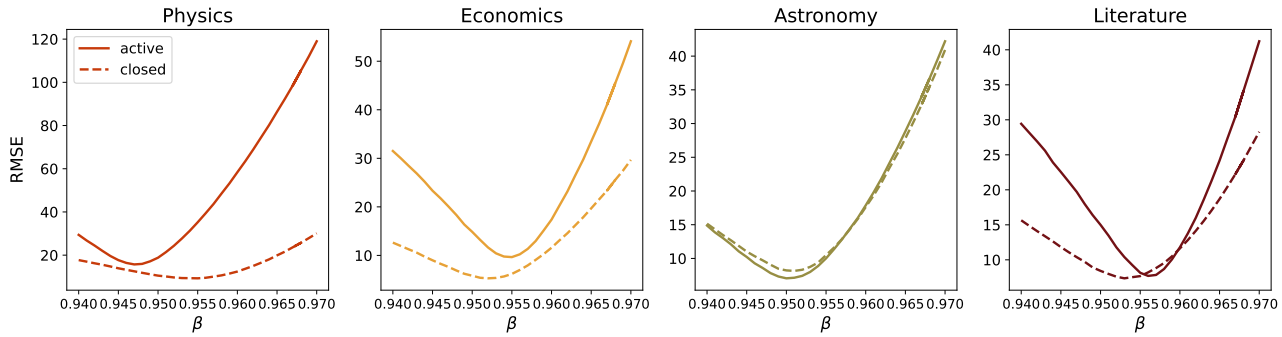

Figure B.2: RMSE between the number of active users in a sliding window of 30 days and the number of users with reputation > 1 for $0.94 < \beta < 0.97$ with step 0.001.

Figure B.3 compares the number of users in the 30-day sliding window and the number of users for these optimal values $\beta = 0.954$ and $\beta = 0.96$. For $\beta = 0.96$, we observe that the estimated number of active users in most communities is consistently slightly higher than the actual number of users who have made at least one interaction in that sliding window. This means that the dynamic reputation model sometimes overestimates the user's reputation, but it is far more important because it never underestimates 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 the model disregards no active users due to the value of the decay parameter.

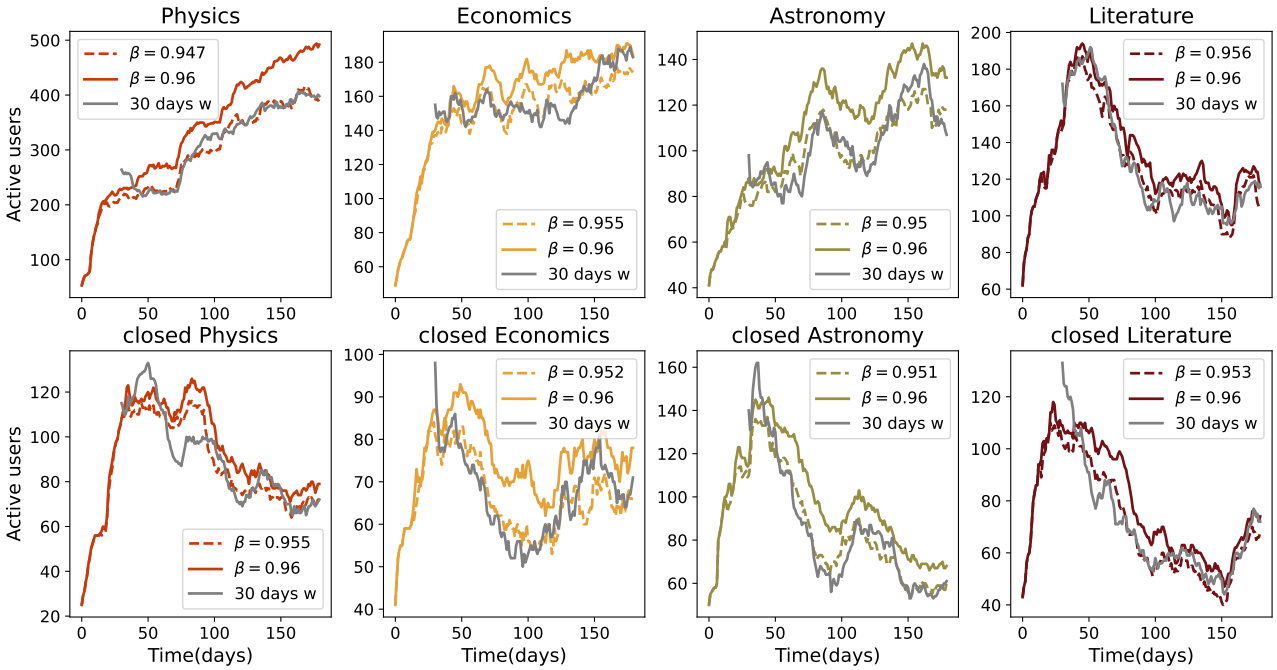

Figure B.3: Number of active users in a sliding window of 30 days and number of users with dynamic reputation higher than 1 for $\beta = 0.954$ and $\beta = 0.96$ which provide the best fit to the number of users in 30 days sub-networks for each community

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

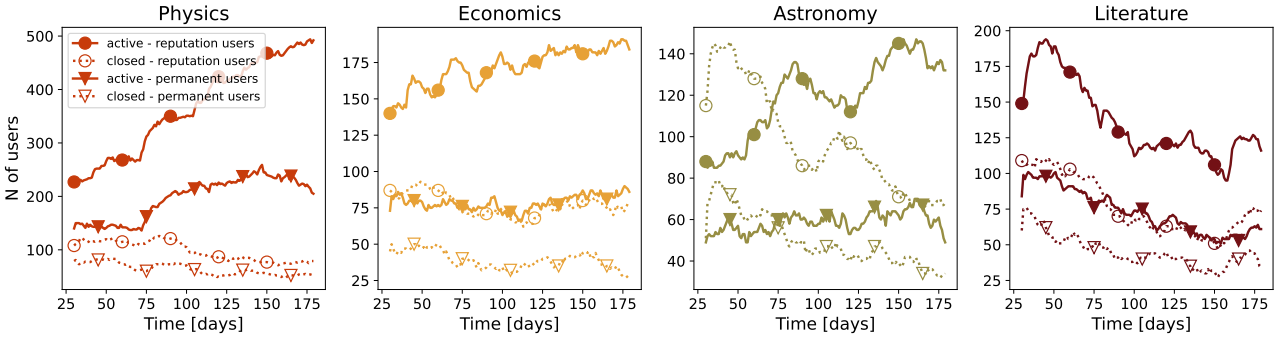

Figure B.4: Solid lines represent the number of users with dynamic reputation higher than 1 for $\beta = 0.96$ while dashed lines are the number of users within 30 days sliding window who were active and remained to be active. Blue lines are beta, while red lines are area51 communities.

Appendix C

The choice of the sliding window

To study the evolution of Stack Exchange communities, we chose to at each time step t analyze the structure of interaction networks created in the period $[t, t + \tau)$. By this, we have better insight into how network properties evolve. However, it is not defined what value the sliding window should take. The previous studies showed that the value of a sliding window determines how much information is saved. If τ is small, sub-networks are sparse, while for a large sliding window, important changes in the measures may not be detected [46, 47]. We analyze how network properties and dynamic reputation depend on the window size. For example, we use Astronomy and compare the active and closed communities, Figure C.1 Similar conclusions can be observed for other pairs of communities. The time window of 30 days approximates one month.

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 apparent that the number of users slightly increases 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. From network measures such as L/N and clustering, we conclude that the difference between closed and active sites is 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, within 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, the number of nodes in the core increases and the results of core-periphery measures and dynamical reputation between core users and between core and periphery users become smoother. Finally, the choice of the sliding window does not change the conclusion that core users in the beta communities produce more activity and make a strong core. However, our main results are shown for a sliding window of 30 days, as it creates a good compromise between large and small time scales.

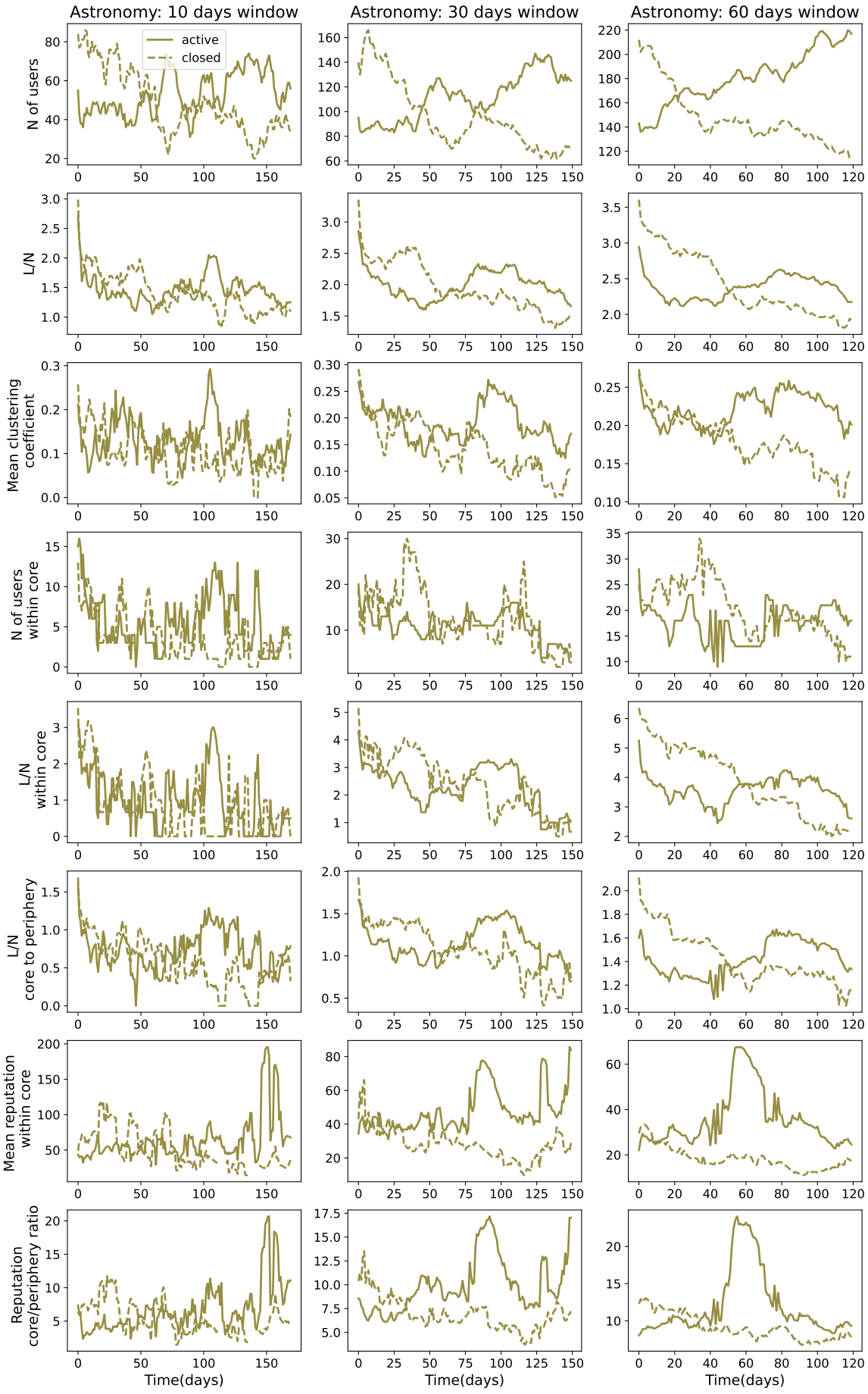

Figure C.1: Results for different sliding windows. For astronomy, solid blue lines- active, orange dashed lines - closed site.

Appendix D

Robustness of core-periphery algorithm

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 the algorithm's robustness, 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 [117] 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) [117].

The **F1 measure** is the harmonic mean of precision and recall [117]:

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

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

The **Jaccard's** coefficient is the ratio of two classes' intersection to their union [117]. It can also be expressed in terms of a confusion matrix:

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

Normalized mutual information (NMI) is similarity measure between to partitions C and $C^{'}$ based on information theory [118]:

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

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.

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 [119] 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 [120], 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.

For example, we show an analysis of an inferred sample of core-periphery structures for 30 days of 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 the similarity between obtained core-periphery configurations in the sample more deeply, we calculate several measures between pairwise 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 the stability of the inferred core-periphery structures.

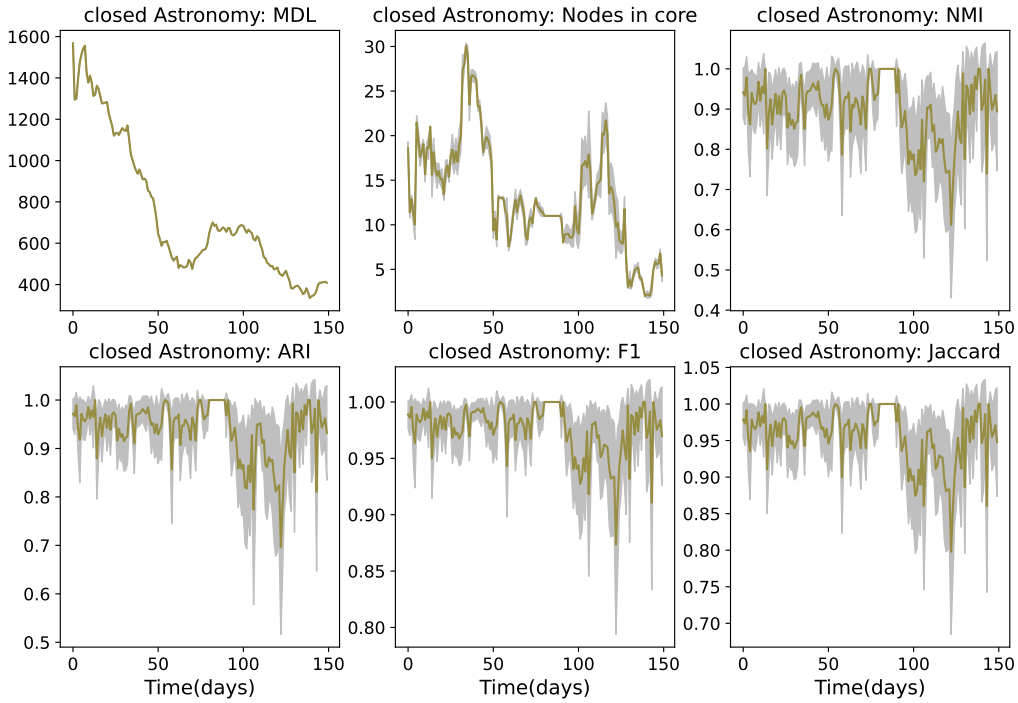

Figure D.1: Minimum description length, number of nodes in the core, normalized mutual information, adjusted rand index, F1 measure and Jaccard index, among 50 samples for 30-days sub-networks. Results are given for closed astronomy.

Bibliography

[1] J. Kwapień and S. Drożdż. Physical approach to complex systems. *Phys. Rep*, 515:115–226, 2012.

[2] Stefan Thurner, Rudolf Hanel, and Peter Klimekl. 93Scaling. In *Introduction to the Theory of Complex Systems*. Oxford University Press, 09 2018.

[3] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The theory of critical phenomena: an introduction to the renormalization group*. Oxford University Press, 1992.

[4] Antonios Garas, David Garcia, Marcin Skowron, and Frank Schweitzer. Emotional persistence in online chatting communities. *Scientific Reports*, 2(1):1–8, 2012.

[5] Santo Fortunato and Claudio Castellano. Scaling and universality in proportional elections. *Physical review letters*, 99(13):138701, 2007.

[6] Arnab Chatterjee, Marija Mitrović, and Santo Fortunato. Universality in voting behavior: an empirical analysis. *Scientific reports*, 3(1):1–9, 2013.

[7] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[8] M. Barthelemy. The statistical physics of cities. *Nat. Rev. Phys*, 1:406–415, 2019.

[9] Giorgio Fazio and Marco Modica. Pareto or log-normal? best fit and truncation in the distribution of all cities. *Journal of Regional Science*, 55(5):736–756, 2015.

[10] Luís A Nunes Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, H Eugene Stanley, and Michael HR Stanley. Scaling behavior in economics: I. empirical results for company growth. *Journal de Physique I*, 7(4):621–633, 1997.

[11] Michael HR Stanley, Luis AN Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, and H Eugene Stanley. Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806, 1996.

[12] V. Verbavatz and M. Barthelemy. The growth equation of cities. *Nature*, 587:397–401, 2020.

[13] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013.

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

[15] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, pages 913–924, 2014.

[16] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3:9, 2014.

[17] D. Fraiman, P. Balenzuela, J. Foss, and D. R. Chialvo. Ising-like dynamics in large-scale functional brain networks. *Phys. Rev. E*, 79:061922, 2009.

[18] C. M. Schneider, L. de Arcangelis, and H. J. Herrmann. Modeling the topology of protein interaction networks. *Phys. Rev. E*, 84:016112, 2011.

[19] 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.

[20] 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.

[21] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45:167–256, 2003.

[22] Bernardo A Huberman and Lada A Adamic. Growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.

[23] S. Dorogovtsev. *Complex networks*. Oxford: Oxford University Press, 2010.

[24] Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.

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

[26] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393:440–442, 1998.

[27] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[28] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Phy. Rev. E*, 62:1842, 2000.

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

[30] Jin Liu, Jian Li, Yadang Chen, Xianyi Chen, Zhili Zhou, Zejun Yang, and Cheng-Jun Zhang. Modeling complex networks with accelerating growth and aging effect. *Physics Letters A*, 383(13):1396–1400, 2019.

[31] T. Pham, P. Sheridan, and H. Shimodaira. Joint estimation of preferential attachment and node fitness in growing complex networks. *Sci. Rep*, 6:32558, 2016.

[32] Parongama Sen. Accelerated growth in outgoing links in evolving networks: Deterministic versus stochastic picture. *Physical Review E*, 69(4):046107, 2004.

[33] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in blog space. *The European Physical Journal B 2009 73:2*, 73(2):293–301, 2009.

[34] Marija Mitrović and Bosiljka Tadić. Emergence and structure of cybercommunities. In *Springer Optimization and Its Applications*, volume 57, pages 209–227. Springer International Publishing, 2012.

[35] 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.

[36] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4628, Nov 2000.

[37] Bosiljka Tadić. Dynamics of directed graphs: The world-wide web. *Physica A: Statistical Mechanics and its Applications*, 293(1-2):273–284, 2001.

[38] Marija Mitrović and Bosiljka Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(2):026123, 2009.

[39] Gourab Ghoshal, Liping Chi, and Albert-László Barabási. Uncovering the role of elementary processes in network evolution. *Scientific reports*, 3(1):1–8, 2013.

[40] 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.

[41] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.

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

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

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

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

[46] 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.

[47] 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.

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

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

[50] 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.

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

[52] 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.

[53] Ana Vranić, Jelena Smiljanić, and Marija Mitrović Dankulov. Universal growth of social groups: empirical analysis and modeling. *arXiv preprint arXiv:2206.06732*, 2022.

[54] Ana Vranić, Aleksandar Tomašević, Aleksandra Alorić, and Marija Mitrović Dankulov. Sustainability of stack exchange q\&a communities: the role of trust. *arXiv preprint arXiv:2205.07745*, 2022.

[55] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *Science advances*, 7(12):eabc9800, 2021.

[56] 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.

[57] Sergey N Dorogovtsev and Jose FF Mendes. Evolution of networks. *Advances in physics*, 51(4):1079–1187, 2002.

[58] Maarten Van Steen. Graph theory and complex networks. *An introduction*, 144, 2010.

[59] Juyong Park and Mark EJ Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.

[60] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.

[61] Mark EJ Newman. Random graphs with clustering. *Physical review letters*, 103(5):058701, 2009.

[62] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.

[63] 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.

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

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

[66] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.

[67] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106, Apr 2010.

[68] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[69] Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 2013.

[70] Fragkiskos D Malliaros, Christos Giatsidis, Apostolos N Papadopoulos, and Michalis Vazirgiannis. The core decomposition of networks: Theory, algorithms and applications. *The VLDB Journal*, 29(1):61–92, 2020.

[71] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social networks*, 21(4):375–395, 2000.

[72] Xiao Zhang, Travis Martin, and Mark EJ Newman. Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803, 2015.

[73] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.

[74] Béla Bollobás and Oliver M Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.

[75] Pavel L Krapivsky, Sidney Redner, and Eli Ben-Naim. *A kinetic view of statistical physics*. Cambridge University Press, 2010.

[76] Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.

[77] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.

[78] Martin Rosvall, Jean-Charles Delvenne, Michael T Schaub, and Renaud Lambiotte. Different approaches to community detection. *Advances in network clustering and blockmodeling*, pages 105–119, 2019.

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

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

[81] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[82] Eckhard Limpert, Werner A Stahel, and Markus Abbt. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience*, 51(5):341–352, 2001.

[83] A.-L. Barabási. Network science book. *Network Science*, 625, 2014.

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

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

[86] Jan W Kantelhardt. Fractal and multifractal time series. *arXiv preprint arXiv:0804.0747*, 2008.

[87] Chao Fan, Jin-Li Guo, and Yi-Long Zha. Fractal analysis on human dynamics of library loans. *Physica A: Statistical Mechanics and its Applications*, 391(24):6617–6625, 2012.

[88] Sergei Sidorov, Alexey Faizliev, and Vladimir Balash. Fractality and multifractality analysis of news sentiments time series. *IAENG International Journal of Applied Mathematics*, 48(1), 2018.

[89] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799, 1951.

[90] Kun Hu, Plamen Ch Ivanov, Zhi Chen, Pedro Carpena, and H Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1):011114, 2001.

[91] Jan W Kantelhardt, Eva Koscielny-Bunde, Henio HA Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3-4):441–454, 2001.

[92] 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.

[93] E. Alexander F. E.A.F.I. Ihlen. Introduction to multifractal detrended fluctuation analysis in Matlab. *Front. Psychol.*, 3:141, 2012.

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

[95] M. Šuvakov, M. Mitrović, V. Gligorijević, and B. Tadić. How the online social networks are used: dialogues-based structure of MySpace. *Journal of The Royal Society Interface*, 10:20120819, 2013.

[96] 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.

[97] 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.

[98] 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.

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

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

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

[107] 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.

[108] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999.

[109] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.

[110] C. Orsini, M. Mitrović Dankulov, P. Colomer-de Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and Kriukov D. Quantifying randomness in real networks. *Nat. Commun*, 6:8627, 2015.

[111] 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.

[112] 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.

[113] 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.

[114] X. Gabaix. Zipf's Law and the Growth of Cities. *Am. Econ. Rev.*, 89:129–132, 1999.

[115] 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.

[116] 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.

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

[118] 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.

[119] 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.

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

Biography of the author

Изјава о ауторству

Име и презиме аутора – **Име Презиме**

Број индекса – **XXXXXXXXX**

Изјављујем

да је докторска дисертација под насловом

Naslov teze na engleskom

(Наслов тезе на српском)

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршила ауторска права и користила интелектуалну својину других лица.

У Београду, 2022 **Потпис аутора**

_____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора – **Име Презиме**

Број индекса – **XXXXXXXXXXXX**

Студијски програм – Физика кондензоване материје и статистичка физика

Наслов рада – **English title of thesis**

(Српски наслов рада)

Ментор – **др Петар Петровић**

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предала ради похрањивања у **Дигиталном репозиторијуму Универзитета у Београду**.

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

У Београду, 2022 **Потпис аутора**

_____

Изјава о коришћењу

Овлашћујем Универзитецку библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

Naslov na engleskom

(Наслов на српском)

која је моје ауторско дело.

Дисертацију са свим прилозима предала сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Цреативе Цоммонс) за коју сам се одлучила.

1. Ауторство (ЦЦ БЫ)
2. Ауторство – некомерцијално (ЦЦ БЫ-НЦ)
3. Ауторство – некомерцијално – без прерада (ЦЦ БЫ-НЦ-НД)
4. **Ауторство – некомерцијално – делити под истим условима (ЦЦ БЫ-НЦ-СА)**
5. Ауторство – без прерада (ЦЦ БЫ-НД)
6. Ауторство – делити под истим условима (ЦЦ БЫ-СА)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

У Београду, 2022 **Потпис аутора**

_____

1. **Ауторство.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима.** Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.