UNIVERSITY OF BELGRADE
FACULTY OF PHYSICS

Ana Vranić

EVOLVING COMPLEX NETWORKS: STRUCTURE AND DYNAMICS

Doctoral Dissertation

Belgrade, 2023

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЗИЧКИ ФАКУЛТЕТ

Ана Вранић

РАСТУЋЕ КОМПЛЕКСНЕ МРЕЖЕ: СТРУКТУРА И ДИНАМИКА

докторска дисертација

Београд, 2023

Thesis Defense Committee

Thesis advisor:

Dr. Marija Mitrović Dankulov
Associate Research Professor
Institute of Physics Belgrade
University of Belgrade

Committee members:

Prof. Dr. Sunčica Elezović Hadžić
Professor
Faculty of Physics
University of Belgrade

Dr. Svetislav Mijatović
Assistant Professor
Faculty of Physics
University of Belgrade

Dr. Antun Balaž
Research Professor
Institute of Physics Belgrade
University of Belgrade

Acknowledgements

This thesis was completed under Dr. Marija Mitrović Dankulov supervision at the Scientific Computing Laboratory at the Institute of Physics Belgrade. I want to express my sincere gratitude to my supervisor for her invaluable guidance, support, and patience during my studies. Her mentorship has been instrumental in helping me to complete my dissertation.

I am grateful to the head of SCL, Dr. Antun Balaz, for his ongoing assistance and advice through all these years. I also want to thank colleagues from the laboratory and institute for making the workplace so enjoyable. I wish to acknowledge collaborators Dr. Aleksandra Alorić, Dr. Jelena Smiljanić, and Dr. Aleksandar Tomasević for their contributions to the research presented in this thesis. Their expertiese, valuable insights, and numerous discussions we had, helped me to refine my research. It has been my pleasure to collaborate with Darja Cvetković and have the opportunity to learn so much from her.

I thank my family and friends for their love and support. To my parents, who gave me tremendous encouragement and understanding, especially to my mom, for being by my side and believing in me.

The research presented here was supported by the Ministry of Education, Science, and Technological Development of the Republic of Serbia, the National Project ON171017 Modeling and Numerical Simulations of Complex Many-Body Systems; by the Science Fund of the Republic of Serbia, the Artificial intelligence theoretical foundations for advanced spatio-temporal modeling of data and processes (ATLAS) project; by Innovation Fund of Republic Serbia the Platform for REmote development of Autonomous Driving algorithms in a realistic environment (READ) project and by 60seconds startup. Numerical simulations were run on the PARADOX supercomputing facility at the Scientific Computing Laboratory of the Institute of Physics Belgrade.

Abstract

Complex systems are all around us and can be found in various domains of physics, biology, and social sciences. While they differ in origin and function, their common feature is that they consist of a large number of interacting units and that due to these interactions exhibit collective behavior. Complex networks represent a general framework for representing interaction patterns in complex systems. The structure of a complex network and its evolution are inevitably linked to the dynamics and function of a complex system. Detecting the collective phenomena and understanding how they emerge from individual interactions is important research problems. Complexity science gives us new ways to explore complex systems. Complexity science combines tools, methods, and paradigms of statistical and computational physics, complex network theory, and computer science to describe and study different collective phenomena quantitatively and propose theoretical models to better understand the mechanisms underlying dynamics and drive the evolution of complex networks.

This thesis aims to broaden the knowledge of the structure and dynamics of evolving complex networks by analyzing the empirical data from different online social systems and providing the models and theories that could explain their specific characteristics. Social systems constantly evolve, and because of that, it is necessary to understand the connections between their structure, growth, and segmentation and how these connections influence their sustainability.

Earlier works have suggested that the properties of growth signals influence the structure and dynamics of evolving complex networks. In real online systems, growth signals fluctuate over time, and they are long-range correlated and have multifractal properties. We use time series of new users from real systems, MySpace and TECH, and computer-generated signals with specific long-range correlation properties as growing signals. We combine them with a network model of aging nodes to examine in detail how the features of these signals shape the structure of complex networks. Our results show that the properties of the growth signal have the substantial influence on the structure of networks with broad degree distribution. Unlike networks grown with constant signals, these networks are clustered and correlated.

Further, we explore the influence of growth signals and linking rules on the segmentation and growth of the social group in the social system. Empirical analysis of different socio-economic systems indicates that despite their differences, these systems often exhibit some universal properties regarding their segmentation and growth. We analyze the Meetup groups in London and New York and Subreddits and find that group size distribution in these systems is lognormal and universal over time, location, and topic. We use a model that interplays two criteria for users' linking with social groups, random and based on social connections. We show that social interactions are an essential factor in the emergence of the lognormal distribution. We demonstrate that mechanisms under which users join social groups could explain the emergence of some universal properties in the social system.

The complex network theory allows us to determine how different network properties evolve and understand how this evolution influences their sustainability. We use data from Stack Exchange sites and compare the evolution of network structure for pairs of active and closed communities during their early phase of existence. Stack Exchange sites are question-and-answer platforms where users share knowledge on some specific topic. We compare active and closed communities on four topics, namely astronomy, literature, economics, and physics. We analyze the structural patterns in these communities and find that active ones are more clustered and characterized by better-connected and stable cores. Core users are crucial for a healthy community and need to be trustworthy. Through the dynamic reputation model, we 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: statistical physics of complex systems, the structure and dynamics of complex networks, modeling online social systems
Research field: Physics
Research subfield: Statistical physics
UDC number: 536

Сажетак

Комплексни системи се налазе свуда око нас у различитим доменима физике, биологије и друштвених наука. Иако се разликују по пореклу и функцији, заједничка карактеристика им је да се састоје од великог броја елемената који међусобно интерагују и због тих интеракција испољавају колективно понашање. Комплексне мреже представљају општи приступ за репрезентацију образаца интеракција у комплексним системима. Структура комплексне мреже и њена еволуција су узајамно повезане са динамиком и функцијом комплексног система. Проналажење колективних фенома и разумевање како они настају из индивидуалних интеракција је један од важних истраживачких проблема. Теорија комплексних система нам пружа нове методе за истраживање комплексних система. Она комбинује методе статистичке физике, рачунарске физике, теорије комплексних мрежа, компјутерских наука како би квантитативно описала и проучавала различите колективне појаве и предложила теоријске моделе ради бољег разумевања механизама који су у основи динамике и еволуције комплексних мрежа.

Ова теза има за циљ да прошири знање о структури и динамици растућих комплексних мрежа кроз анализу емпиријских података из различитих онлајн друштвених система и дефинисањем модела и теорија које би могле да објасне њихове специфичне карактеристике. Друштвени системи стално еволуирају и због тога је неопходно разумети везе између њихове структуре, раста и сегментације и како те везе утичу на њихову одрживост.

Ранији радови сугерисали су да својства сигнала раста утичу на структуру и динамику растућих комплексних мрежа. У реалним онлајн системима, сигнали раста флуктуирају током времена и они су дугодометно корелисани и имају мултифрактална својства. Као сигнале раста у овој тези, користимо временске серије нових корисника из реалних система MyS-pace и TECH, и компјутерски генерисане сигнале са специфичним својствима дугодометних корелација. Комбинујемо их са мрежним моделом старости чворова да бисмо детаљно испитали како карактеристике ових сигнала утичу на структуру комплесних мрежа. Наши резултати показују да својства сигнала раста имају најзначајнији утицај на структуру мрежа са широким степеном дистрибуције. За разлику од мрежа које имају константан раст, ове мреже су кластерисане и корелиране.

Даље, истражујемо како сигнал раста и правила повезивања утичу на сегментацију и раст социјалних група. Емпиријска анализа различитих друштвено-економских система указује на то да упркос разликама, ови системи често испољавају нека универзална својства у погледу сегментације и раста. Проучавали смо Meetup групе настале у Лондону и Њујорку, као и subReddit и открили да је дистрибуција величине група у овим системима логнормална и универзална током времена, не зависи од локације и теме групе. Користили смо модел који комбинује два критеријума за повезивање корисника са друштвеним групама, насумично или

на основу друштвених веза. Показали смо да су друштвене интеракције битан фактор при настанку логнормалне дистрибуције. Механизми под којима се корисници придружују друштвеним групама могу објаснити појаву универзалних својстава у друштвеном систему.

Комплексна теорија мрежа нам омогућава да опишемо како се развијају различита својства мреже и разумемо како еволуција утиче на њихову одрживост. Користили смо податке са Stack Exchange сајтова и упоређивали еволуцију структуре мреже за парове активних и затворених заједница током њихове ране фазе постојања. Stack Exchange сајтови су платформа за питања и одговоре на којима корисници деле знање о некој специфичној теми. Упоредили смо активне и затворене заједнице на четири теме, а то су астрономија, књижевност, економија и физика. Анализирали смо структурне обрасце у овим заједницама и открили да су активне више кластерисане и да их карактерише боље повезана и стабилност језгра. Кроз динамички модел репутације измерили смо ниво поверења у овим заједницама. У активним заједницама, корисници који се налазе у језгру имају већу репутацију него у затвореним заједницама, што указује на важност да се стабилно језгро развије рано и да има висок ниво поверења.

Кључне речи: статистичка физика комплексних система, структура и динамика комплексних мрежа, моделовање онлајн социјалних система
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 organizations, cities, or cells, consist of many interacting units and belong to a class commonly known as complex systems. One of the most prominent characteristics of complex systems is that they exhibit emergent collective behavior that can not be predicted based on the behavior of individual components. The interactions between system components can be represented as a complex network [1]. The emergence of collective behavior strongly depends on the structure of the network of interactions. 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 civilization, economy, and formation of social groups [2]. Also, the animal populations show different levels of organization: such as patterns in bird flocks or schools of fish [2].

Despite the differences between complex systems, they can be studied using the same techniques. The natural extension of the complex system is the network, which consists of sets of nodes (vertices) and links (edges). Elements in the system are nodes, while interactions between them are represented as edges. This approximation allows us to equally approach social [3, 4] (graph of actors), biological (network of proteins) [5, 6] or even technological systems (internet, traffic) [7, 8, 9]. The research in complex systems mainly focuses on the interactions between its units. Knowing the structure of these connections, we can determine the properties of the system [10]. We can construct a representation with neurons and synapses representing connectivity in the brain network [11]. 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.

While the relationships between individuals characterize the structure of complex networks, the dynamics describe changes in individual behaviors over time. As real complex networks constantly evolve, the interactions between their elements can also change [2]. Networks can exhibit the addition of new nodes, removal of existing nodes, or change in the number of edges and the strength in these edges. While these changes occur, the structure, but also the function of the network could be affected. The formation of clusters, hubs, and node removal directly influence network connectivity, and robustness [12].

The application of principles of statistical physics and complex network theory in the study of social systems lead to the creation of the new, interdisciplinary field of socio-physics [13]. It provides methods for the statistical description of the structure and dynamics of social networks. Social

networks are very dynamic, and despite their constant evolution, they show universal properties [14].

Broadly, universality is an important property of complex systems [15]. One of the well-known examples of universality in physics is a phase transition, such as in the Ising model of magnetization [16]. At a critical transition point, the system's properties are independent of the specific details of the system. In the Ising model, a critical point is a temperature at which the system undergoes the phase transition from a disordered phase to an ordered phase. The correlation length of the system diverges and exhibits the power-law scaling. The critical exponents, which describe the scaling of different quantities near the critical point, are the same for the model with different interaction patterns [17]. We also find universal behavior in systems where elements are ordered randomly, as in complex networks. For example, the time lap between two email messages follows the power-law distribution [18], and the exponent is universal across different platforms. Similar conclusions are found in distributions of the votes in elections [19, 20], and citations of scientific publications [21]. 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 [22, 23]. However, the distribution of company sizes follows log-normal behavior and remains stable over decades [24, 25]. Identifying universal behavior and understanding its emergence in the system is one of the main topics in the statistical physics of complex networks [26]. In this thesis, we will explore the structural and dynamical properties of evolving online social networks and apply complex network models.

When constructing complex network models, the specific mechanisms that govern social interaction and lead to observed macroscopic properties in empirical networks must be considered [13]. Many studies confirmed that networks show power-law scaling in the distribution of the number of connections, high clustering, and nodes tend to connect to structurally similar nodes. For that reason, complex network models have been created to mimic properties found in real social systems [13].

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.

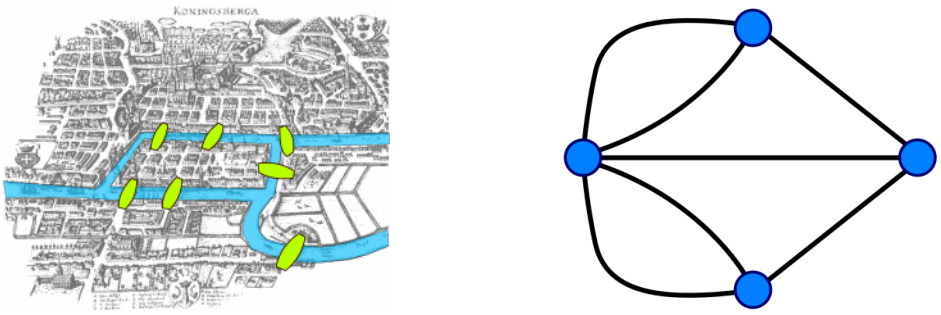

Figure 1.1: The Konigsberg problem of seven bridges. The left panel shows the original map of the bridges; the right panel shows its graph representation.

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 analyze its properties [27], they found that degree distribution follows the power-law contrary to expected behavior from random graph model

[28]. 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 [29, 30]. Two seminal papers from 1999 inspired further research in complex networks. Watts and Strogatz [31] proposed the model where rewiring of edges on regular lattice leads to the network in which paths between any two nodes become short (small-world) and nodes become densely connected, resulting in a high clustering coefficient. On the other hand, Barabasi and Albert (BA) [32] 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 [33, 34]. 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 [35]. Some models considered that nodes become inactive or even that network grows through a nonlinear number of links [36]. On the other hand, models with accelerated growth in the number of nodes [37] 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 behavior [38, 39, 40].

Research has also been devoted to using generated networks to analyze dynamic processes on top of them. Central questions are about the spread of epidemics, information diffusion, or emotional interactions among elements [18]. These systems are modeled 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 [41]. They also show resilience to random attacks. Real-world networks are often characterized 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 [32, 42, 43, 44]. 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 modeling 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 [45, 46, 47]. For example, they are used to determine missing links and recommend to users in online social networks [48, 49] or to develop generative GNN models that lead to the discovery of new drugs [50, 51].

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 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, numerical representation allows us to characterize the statistical properties of the networks. The graph G, with N nodes could be represented with **adjacency matrix** $|A| = N \times N$ [12]. The matrix elements are equal to 1 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.

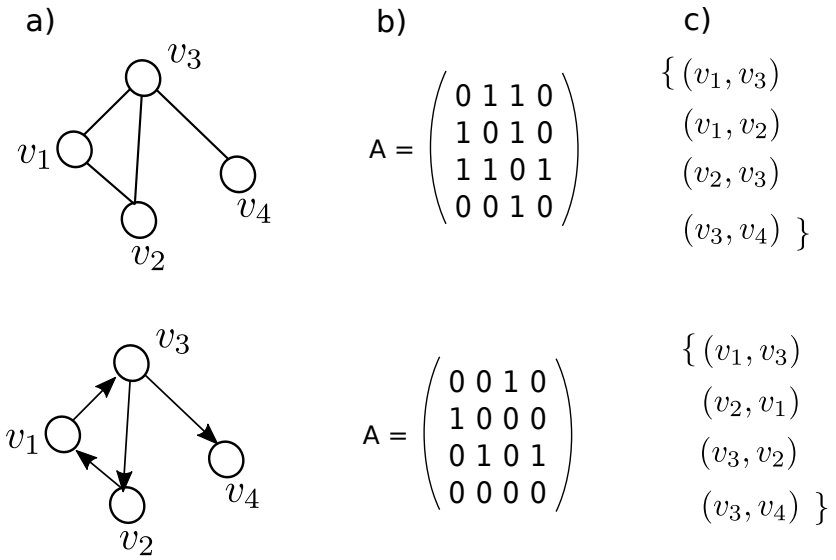

Figure 1.2: a) Graph representation of undirected (top panel) and directed (bottom panel) network. The same networks are represented with adjacency matrices in column b) and edge list representation in column c).

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 [52]. 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 [53]. 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.

Sometimes is essential to include the specific properties of the system in the network representation. For example, to emphasize the frequent interactions between nodes, edges can be assigned with different values; such networks are **weighted**. In a collaboration network, authors who collaborate more often have stronger interaction. 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 characterized by the time when the interaction between nodes happens, which includes the **temporal** component in the network representation, as in phone calls networks. Finally, if two nodes interact differently, the **multigraph** is an appropriate configuration where multiple edges are allowed. The transportation network, consisting of roads and railways, could be seen as a multigraph. Figure 1.3 presents the graphical representation of discussed network representations.

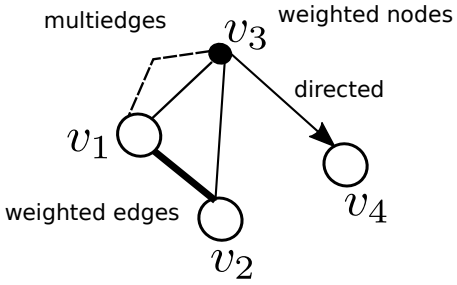

Figure 1.3: The complex networks may represent different system characteristics. The edges can be directed, weighted or multiply. Also, nodes can be assigned with different weights or any relevant feature.

A **bipartite network** consists of two types of nodes. The nodes in the same partition are not connected, while links exist only between partitions, Figure 1.4. For many real systems, a bipartite graph is a natural representation [53, 11]. For example, the bipartite network of people and groups has two distinct node partitions, where links indicate the memberships. Another example is a system of customers and products. The user and item link is created when the user bought an item. The bipartite networks find their application in the algorithms for recommender systems, whose goal is to suggest items that may interest the user. They are often used to find the most probable missing links in the network.

Though the nodes in the same partition of a bipartite network are not directly connected, we can analyze their connections by projecting the bipartite network to one partition. The primary assumption is that two nodes in one partition could be connected if they point to the same node in the other partition. Figure 1.4 shows two projections of the bipartite networks. 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 [30, 53].

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 in how many movies two actors

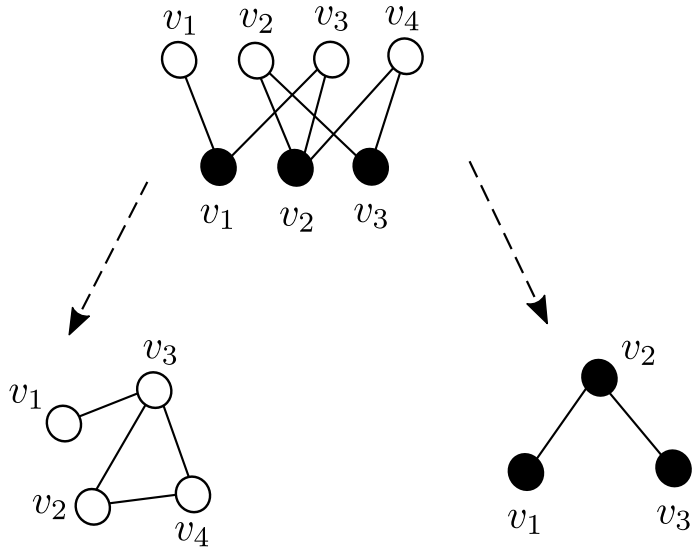

Figure 1.4: Bipartite network and two partition projections.

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-partite 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 [54]. 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 might not be connected, as is the case in the static network.

The temporal network is a collection of timestamped edges; as seen on Figure 1.5 - top panel. 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 [55]. 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.

The temporal network can be represented as a sequence of static networks that evolve in time, $G = \{G(t_1), G(t_2), ..., G(t_{max})\}$, as shown in Figure 1.5 - bottom panel. At each time step, we can create the network and analyze the macroscopic properties of the given network snapshot. With this, we can end up with graph snapshots with many disconnected components or empty graphs for some points [56]. Sometimes, a better approach is aggregating the links 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 $1 \leq \delta t < w$. 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

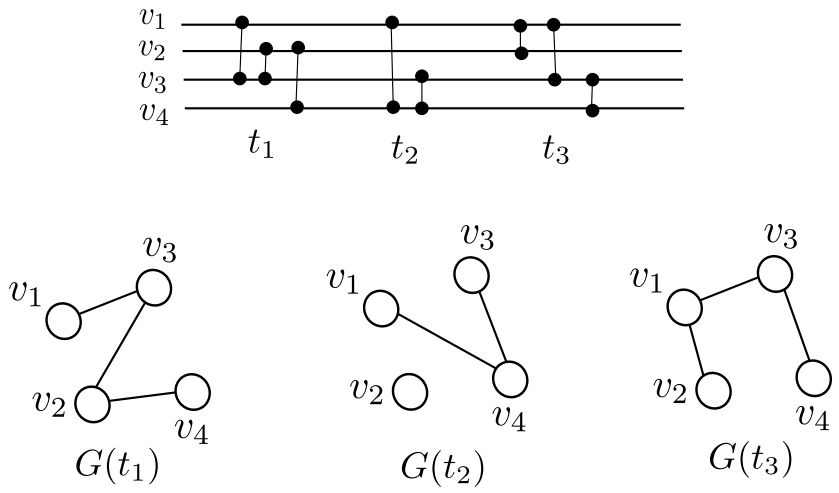

Figure 1.5: Top panel represents temporal network as collection of timestamped edges. Bottom panel represents sequence of static networks.

data are no more available [57, 58].

Multilayer networks were introduced for studying systems in which different types of interaction exist. This formalism allows one to investigate diverse network systems and combine different data types into one model [59]. 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 [60]. Another example is the airline transportation network, where each layer represents the flights of different airline companies [61].

1.2 Thesis outline

This thesis uses combined approaches of statistical physics and complex network theory to model and analyze evolving online social systems. These systems consist of many users interacting online and could be represented by complex networks. The main focus of the thesis is to explore the evolution of these complex networks and understand how different dynamical processes shape their structure. We study the growth of various online social networks using data from Meetup, Reddit, and StackExchange platforms and detect important structural changes in these systems, as well as the processes that lead to the creation of groups and factors important for the emergence of sustainable communities.

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 characterizing 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 the 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 aging model [62], 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

[63], 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 organize; 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 [64]. 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 behavior 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 behavior, 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 [65]. Here, we study the question-answer platform Stack Exchange; it has more than 200 different topic-specific 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 analyze 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 [66]. The core must select firmly connected users, but their interaction with the periphery has to be high. In other words, a trustworthy core is needed to hold the community. Introducing the Dynamical Reputation Model (DIBRM) [67], based on user interaction sequences, we quantify how much users can be trusted and whether a community has a strong core. We briefly describe the Stack Exchange sites in the appendix A. 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 network structure

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, $E = \{e_1, e_2...e_L\}$. There are a lot of measures to quantify the structure of the network. This section describes some of the important measures and their definitions on the undirected and unweighted networks, where the **adjacency matrix** $A = N \times N$ has value 1 if there is a connection between two nodes; otherwise, it is 0 [12]; as this network representation is mostly used through the thesis. We list 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 is the number of nodes adjacent to node i, $k_i = \sum_j A_{ij}$ [12, 30]. The network density is the average degree divided by $N - 1$, where N is the number of nodes [68].

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 [69].

The degree distribution is the probability, $P(k)$, that a randomly chosen node has degree k [30, 68]. To estimate the degree distribution, we can consider the fraction of k degree nodes N_k, $p(k) = N_k/N$. Similarly, we can order nodes according to their degree and plot the node degree.

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 [53].

- Exponential distribution. $P(k) = e^{-k/k}$. It is the degree distribution of the growing random graph [53]. 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 [53, 30]. $P(k) = k^{-\gamma}$, 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 [30].

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 [52]. 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-degree correlations

Correlation is defined through a correlation coefficient $r(x, y)$. For two variables x and y, which represent pairs $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ we can define correlation coefficient [70] as:

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

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 [70]. For graph G which consists of n nodes and is characterized with with adjacency matrix $\boldsymbol{A}$ and degree sequence $\boldsymbol{d} = [d_1, ..., d_n]$, correlation of vertex degree has form:

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

An adjacency matrix allows us to calculate the correlations between neighboring 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 index**. If correlations are positive, networks are assortative; there is a tendency for connections to exist between similar degree nodes [53]. The negative correlations indicate that nodes with large degree are more likely to connect nodes with small degree, disassortative networks. The average first neighbor degree k_{nn} can be calculated as $k_{nn} = \sum_{k'} k' P(k'|k)$. The P is 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 [12], $kP(k'|k)P(k) = k'P(k|k')P(k')$ [12]. 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 [71].

The Newman defined the assortativity [72] 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 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 neighborhood's structure. In networks, exist a tendency to form triangles or clusters [53]. It is common property of friendship networks; there is high probability that neighbors of one nodes are connected [73]. The clustering of node i can be measured as [31]:

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

where e_i is number of links among neighbors of node i and k_i is node degree.

We can calculate the mean clustering coefficient by averaging it overall network nodes. It ranges from $\langle c \rangle = 0$ where connections between neighboring nodes do not exist; the network has a tree structure [53]. On the other hand, $\langle c \rangle = 1$ indicates a fully connected network [53].

Alternative definition of the clustering coefficient was proposed by Newman [74]. The network transitivity is seen as global clustering as it takes into account whole network properties. It is calculated as ratio of number of triangles and triples in the network. While triangle is complete subgraph of tree nodes, a triple has tree nodes, but only two edges.

2.1.4 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 [70], 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 [75].

We can find different paths between two nodes in the network, but the most important one is the **shortest path** [70, 75]. 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, but its length is not necessary 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 [70]. These definitions apply to directed and undirected graphs.

Also for each node u in network G we can calculate the average length of the shortest paths to any other node in the network [70]:

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

The **average path length** of the network is then calculated as:

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

while it is also possible to define the **characteristic path** length of G as median value of all nodes shortest paths.

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 the 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 [76]:

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

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. Parameter $0 \leq \omega \leq 1$ determines importance of first and second term in D-measure. 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 [77, 78], 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 [79] 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.

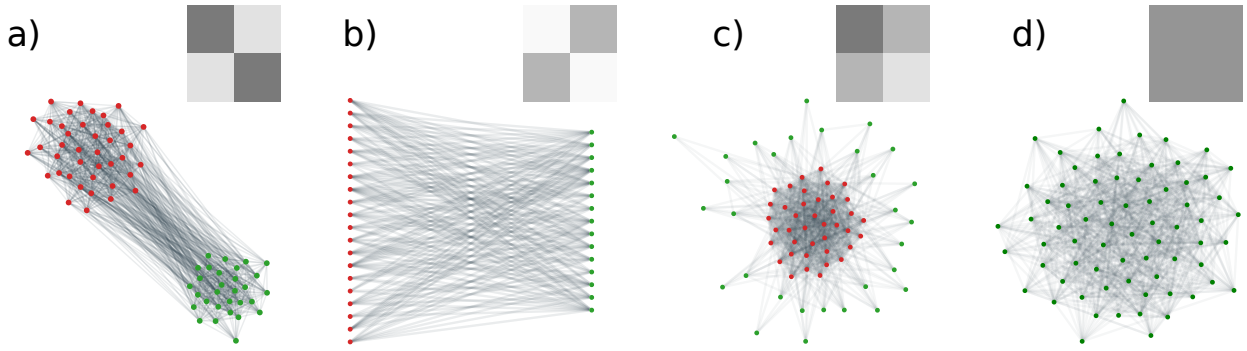

Figure 2.1: Different communities structures (a) assortative. (b) disassortative. (c) core-periphery. (d) Erdos Renyi random graph.

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 [80]. Even though community detection algorithms are widely used in complex network analysis as they can give us a better understanding of network structure [80, 81]. In this section are explained two community detection models, the first one based on optimizing the modularity function [77, 82], and the other based on the statistical inference of the Stochastic Block Model (SBM) where is optimized the likelihood function [77, 83, 84].

2.2.1 Community detection based on modularity function optimization

The **modularity** [85, 82, 86] is a measure used to evaluate the quality of a partition or clustering of nodes into communities. Partition is the division of the network with N nodes and L links into n_c communities, where each node belongs to only one group [87]. The modularity measures the degree to which nodes in the same community are more connected to each other than expected by chance, while taking into account the expected degree sequence of the network. The modularity has form:

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

where the first part of equation measures number of links A_{ij} within community c, while second term is number of links within community if network is randomly connected $p_{ij} = \frac{k_i k_j}{2L}$. If the first term is larger than the second term, the modularity is positive and the partition is considered to be better than random, otherwise we can not consider that nodes in given group form community structure. 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. The Newman showed that modularity function [88] applies for weighted networks.

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 we get single community. The best partition is one with the largest M.

Louvain algorithm [89] is an optimization algorithm with better scalability than the greedy algorithm so it can operate on very large networks. Initially, each node is in different communityand similar to before, we calculate the difference in the modularity moving nodes to one of their neighboring community. Then we move node i to the community such that modularity becomes larger. It is repeated with all nodes in the network, until there is no improvement in the modularity. 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 [87]. The number of links inside the community is given as a weighted self-loop. Then, the first and second steps are repeated until there is no more change in the modularity. The obtained number of clusters when the algorithm stops is an optimal number of communities.

The community detection algorithms tend to merge small communities, which should be independent [90]. This consequence is easily seen in the graph consisting of N-connected cliques, where higher modularity is if two adjacent cliques are merged into communities instead of having each clique as a single community. This lead to the modification of the modularity function as $M = \frac{1}{2L}\sum_{i,j}[A_{i,j} - \gamma\frac{k_i k_j}{2L}]$, where γ is resolution parameter [91], which controls the size of communities to be detected. With $\gamma < 1$, detecting small communities undetected with the original model would be possible.

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 depend on the probabilities for group memberships [83]. In one group, nodes have similar connectivity patterns. To describe the network $G(N, L)$ 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} is

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

The benefit of this model is that we can **generate** many networks with similar network structure [92]. 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)** [84]. The variable which has probability P(x), is described with amount of information $-log_2 P(x)$. 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). 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 decreases with the number of groups [84]. The second contribution comes only from the model, and as the model becomes more complex, with a larger number of groups, this part increases [84]. 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 [93]. In some social networks, users can belong to more than one group, which can be modeled with mixed membership SBM. Other extensions include application to bipartite, weighted network, and hierarchical model [94]. 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

The core-periphery structure is characterized by a group of densely connected nodes in the core, which are more connected to each other than to the less connected nodes in the periphery [95, 30]. The condition $p_{11} > p_{12} > p_{22}$ implies that the probability of edges within the core is higher than the probability of edges between the core and the periphery, which in turn is higher than the probability of edges within the periphery. One way to identify the core-periphery structure is to use the degree criterion, which assumes that the core nodes have higher degrees in the core than in the periphery. Another approach is to use k-cores [96], which are groups of nodes that are connected to at least k other members of the group. The k-cores form a nested hierarchy, and the core-periphery structure can be detected by identifying the densest k-core. Borgatti and Everett [97] proposed a measure similar to modularity to detect core-periphery structures, where the goal is to minimize the number of edges in the periphery. The score function ρ balances the number of observed edges in the periphery with the expected number of edges in a null model where the nodes in the periphery are randomly connected. The optimization problem seeks to maximize the score function ρ, which is defined as $\rho = \frac{1}{2}\sum_{ij}(A_{ij} - p)g_i g_j$, where A_{ij} is the adjacency matrix of the network, p is the expected probability of an edge between two nodes, and g_i is a variable that indicates whether node i belongs to the core or the periphery.

Another way to detect core-periphery structure is to use the inference method based on fits to a Stochastic Block Model (SBM) [98, 93]. 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 [66]. 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 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 [53].

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 [53]. 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. Figure 2.2 shows how different distributions look on linear (first column) and log-log scale (second column).

2.3.1 The properties of distributions

Power-law distribution. The power-law distribution [99, 100] is defined as

$$p(k) = Ck^{-\gamma}, \tag{2.18}$$

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 [52]. We'll find that these criteria are not satisfied by any other distribution

$$\frac{p(k)}{p(2k)} = \frac{Ak^{-\gamma}}{A(2k)^{-\gamma}} = 2^{\gamma}, \tag{2.19}$$

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 [101]

$$f(y) = \frac{1}{2\pi\sigma}e^{-(y-\mu)^2/2\sigma^2}, \tag{2.20}$$

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

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}e^{-(log(x)-\mu)^2/2\sigma^2}. \tag{2.21}$$

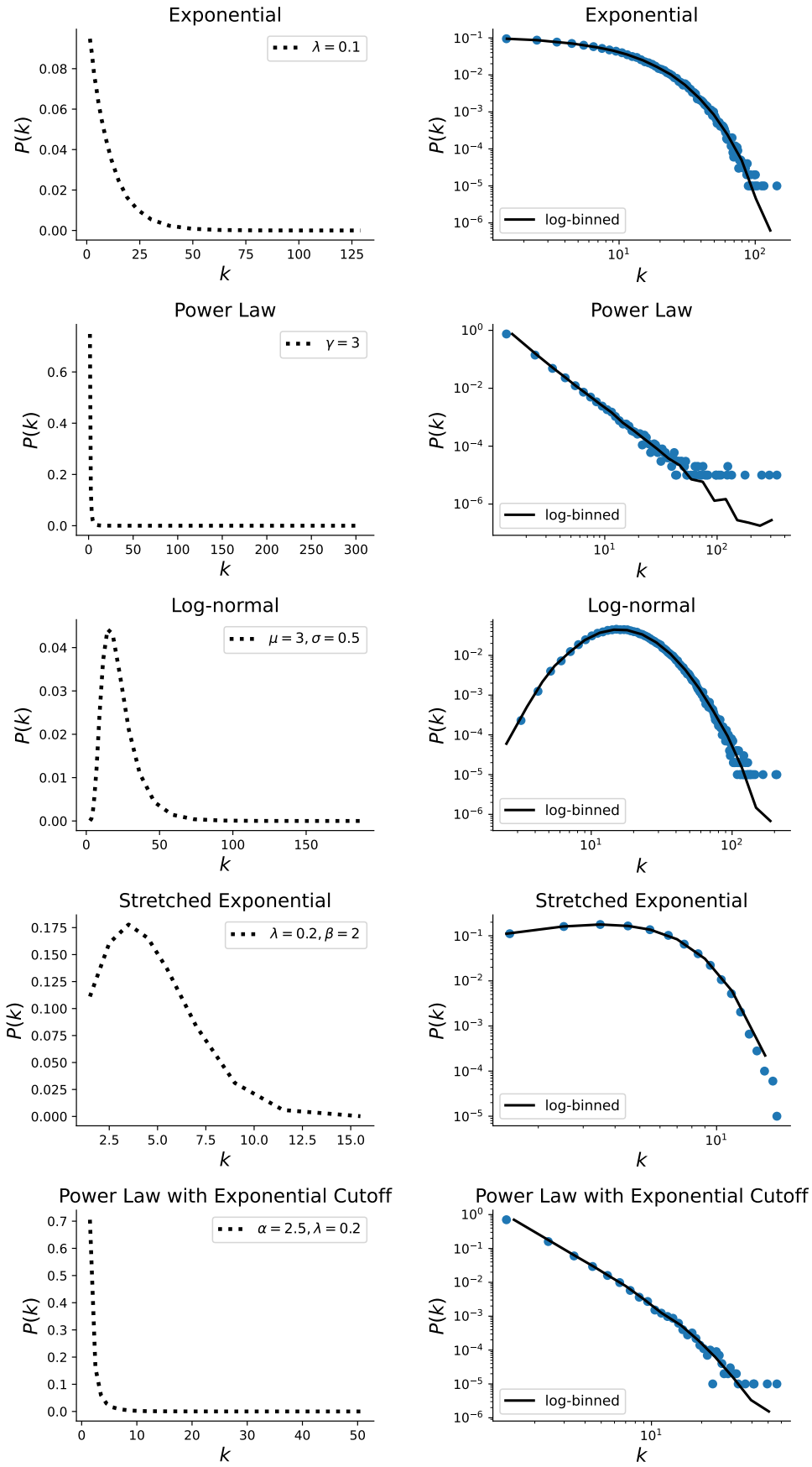

Figure 2.2: Probability distributions on a linear and double logarithmic scale.

The lognormal distribution has finite mean $e^{\mu+1/2\sigma^2}$, and the variance $e^{2\mu+\sigma^2}(e^{\sigma^2-1})$. [99]. 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 [52, 99]. 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 [99]

$$S_t = \epsilon_t S_{t-1}. \tag{2.22}$$

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.:

$$S_t = \epsilon_t S_{t-1} = \epsilon_t \epsilon_{t-1} ... \epsilon_2 \epsilon_1 S_0. \tag{2.23}$$

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:

$$ln(S_t) = ln(S_0) + \sum_{i=0}^{t} ln(\epsilon_i). \tag{2.24}$$

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 [99].

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 [52], 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 is 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

$$p(k) = Ck^{-\gamma} e^{-\lambda k}. \tag{2.25}$$

where $k > 0$ and $\gamma > 0$. This function combines the power-law and exponential terms responsible for an exponentially bounded tail [53]. 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:

$$p(k) = ck^{\beta-1} e^{-(\lambda k)^\beta}. \tag{2.26}$$

the parameter β is stretching exponent determining the properties of the function $p(k)$ [53]. 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.

2.3.2 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 [102]. The likelihood function tells us how likely it is to have the given data if the distribution parameters are θ

$$L(\theta|x_1, ...x_n) = \prod_{i=1}^{i=n} p(x_i|\theta). \tag{2.27}$$

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

First, we fit empirical data to get model parameters and calculate the KS statistics of this fit [103]. 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 [103]. 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

$$\mathcal{R} = \sum_{i=1}^{n} \left[logp_1(x_i) - logp_2(x_i) \right]. \tag{2.28}$$

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

$$p = \frac{1}{\sqrt{2\pi n\sigma^2}}\int_{-\infty}^{-|\mathcal{R}|} e^{-x^2/2n\sigma^2}dx + \int_{|\mathcal{R}|}^{\infty} e^{-x^2/2n\sigma^2}dx. \tag{2.29}$$

Here we use the standard two-tail hypothesis test [103], 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 favor 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 $R = 0$.

2.4 Network models

The interest in analyzing real-world networks allowed us to describe their statistical properties and formulate models to explain essential data features. With network models, we can understand the origins of the properties of complex networks, what mechanisms influence the generation of the network, and how network properties emerge [30, 53]. This section considers the random network and small-world models, which are static models, as the number of nodes is fixed. Even though the random network model is not applicable to real networks, it is important historically as one of the first network models. The small-world model explains how properties of real networks, such as high clustering and small distances may emerge. On the other hand, generative models, such as models of preferential attachment, where the network grows according to specific growing rules, are important for understanding how network structure is created. They allow us to explore different growing mechanisms, and by comparing obtained networks with real data, we can conclude which growth processes have an influence on the network structure.

2.4.1 Random network model

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

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

- We start with N isolated nodes.

- Between each $N(N-1)/2$ pair of nodes we create link with probability p; sampling random number $r \in (0,1)$, we create link if $r \leq p$, see Figure 2.3.

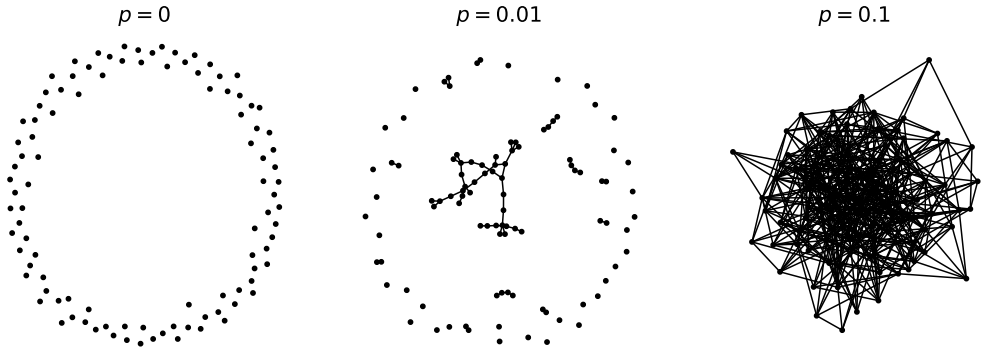

Figure 2.3: Erdős-Rényi graph with $N = 100$ nodes and different linking probabilities p.

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

$$\langle L \rangle = \frac{N(N-1)}{2} p. \tag{2.30}$$

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:

$$\langle k \rangle = (N-1)p. \tag{2.31}$$

The **degree distribution** of ER random graph follows the binomial distribution [53].

$$P(k) = \binom{N-1}{k} p^k (1-p)^{N-1-k}. \tag{2.32}$$

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, see Figure 2.4. 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, as could be shown in Figure 2.4, which now depends only on one parameter $\langle k \rangle$

$$p(k) = \frac{1}{k!} e^{-\langle k \rangle} \langle k \rangle^k. \tag{2.33}$$

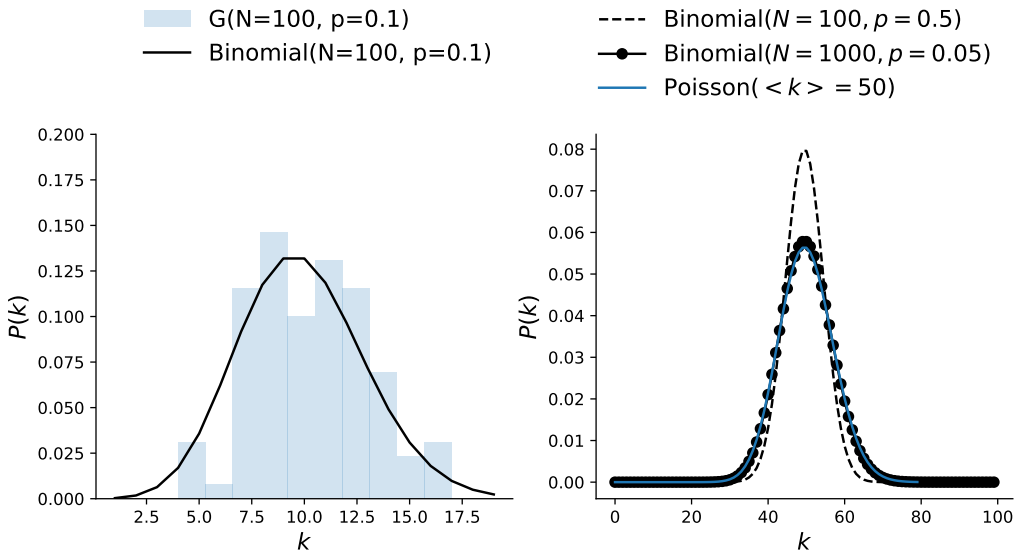

Figure 2.4: Degree distribution of ER graph. The degree distribution of small networks follows binomial. Larger networks are better approximated with Poison distribution, and degree distribution for fixed average degree $< k >$ becomes independent of the network size.

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 [104]. 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 2.3 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 [53]. 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.

2.4.2 Small-world networks

Inspired by the idea that real-world networks are highly clustered and the average distance is small, Watts and Strogatz [31] 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, see Figure 2.5, and each node is connected to $k/2$ first neighbors 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 connects long-distance nodes, decreasing the network's average path length, Figure 2.5.

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 relatively small probabilities ranging from $p \approx 0.01 - 0.1$ [31].

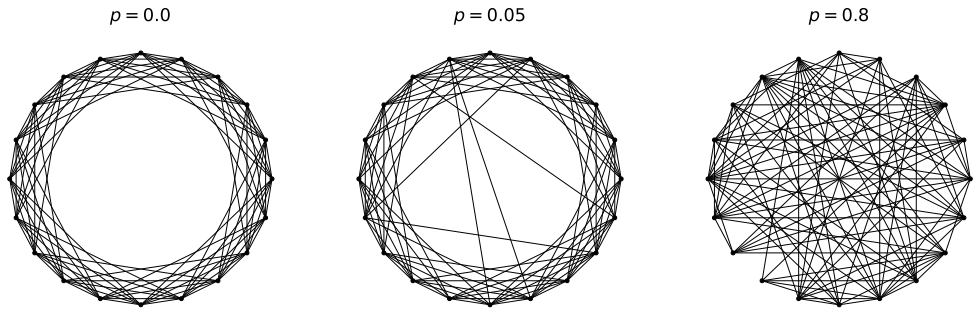

Figure 2.5: Watts and Strogatz graph model creation, for different rewiring probabilities.

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.

2.4.3 Barabási-Albert model

The ER random graph model and WS small-world model are static models where the number of nodes is fixed. It is one of the reasons why they can not fully explain the properties of real systems. The size of real systems does not remain constant; real networks grow. 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 [32].

- 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 [105]

$$\Pi(k_i) = \frac{k_i}{\sum_j k_j}. \tag{2.34}$$

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 [105].

$$\frac{dk_i}{dt} = m\Pi(k_i) = m\frac{k_i}{\sum_j k_j} = m\frac{k_i}{m_0 + 2mt}. \tag{2.35}$$

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

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}. \tag{2.36}$$

With this equation, we can calculate the probability that node has a degree smaller than k [105] as $P[k_i(t) < k] = P(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}})$. Assuming that we add nodes in constant time intervals, we have $P(t_i) = 1/(m_0 + t)$. The cumulative probability is then $P(t_i > \frac{m^{1/\beta}t}{k^{1/\beta}}) = 1 - \frac{t}{t+m_0}\left(\frac{m}{k}\right)^{1/\beta}$. Finally, the degree distribution has the following form

$$P(k) = \frac{\partial P[k_i(t) < k]}{\partial k} \sim 2m^2k^{-3}. \tag{2.37}$$

Degree distribution follows power-law, and for large k is approximated with $P(k) = k^{-\gamma}$, where $\gamma = 3$. As the network grows, nodes with larger degrees become bigger, and we end up with few nodes with many links, called hubs. Figure 2.6 - left pane shows generated BA network, consisting of $N = 100$ nodes, where even on this scale, we can notice the emergence of hubs. The right pane of Figure 2.6 shows obtained degree distribution of a larger network with $N = 10^4$ nodes. The degree distribution is also independent of the time and size of the system, meaning the emergence of a stationary scale-free state. 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 [53].

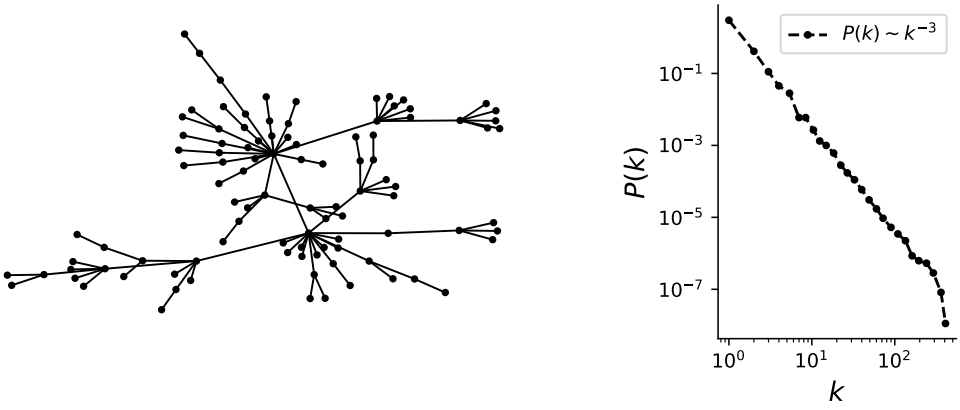

Figure 2.6: Barabasi-Alber model. The left panel shows the BA network, with 100 nodes. The right panel shows the degree distribution for BA network of 10^4 nodes that follow the power-law.

The **network diameter**, represents the maximum distance in network, $d \sim \frac{lnN}{lnlnN}$ [104]. 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 $C \sim \frac{lnN^2}{N}$ [104]. 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 [32]. 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.

2.4.4 Nonlinear preferential attachment model

In the nonlinear preferential attachment model linking probability also depends on the node degree. The dependence is not linear and has the following a form [106]:

$$\Pi(k_i) = {k_i}^{\beta}. \tag{2.38}$$

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 [106].

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.

2.4.5 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 [33]

$$\Pi_i(t) \sim k_i \tau_i^{\alpha}. \tag{2.39}$$

The parameter α controls the linking probability dependence on the nodes' age, as could be seen on Figure 2.7. If $\alpha = 0$, the aging 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, aging 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 aging 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 [62]

$$\Pi_i(t) \sim k_i(t)^{\beta} \tau_i^{\alpha}. \tag{2.40}$$

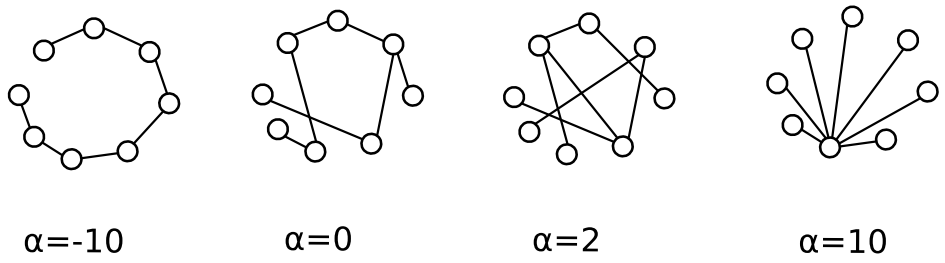

Figure 2.7: Dependence of parameter α and network structure. Network topology vary from chain network to the case where each node is connected to youngest node.

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.
- Scale-free property can be produced along the critical line $\beta(\alpha^*)$ in the $\alpha - \beta$ phase diagram, see Figure 2.8.
- 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.

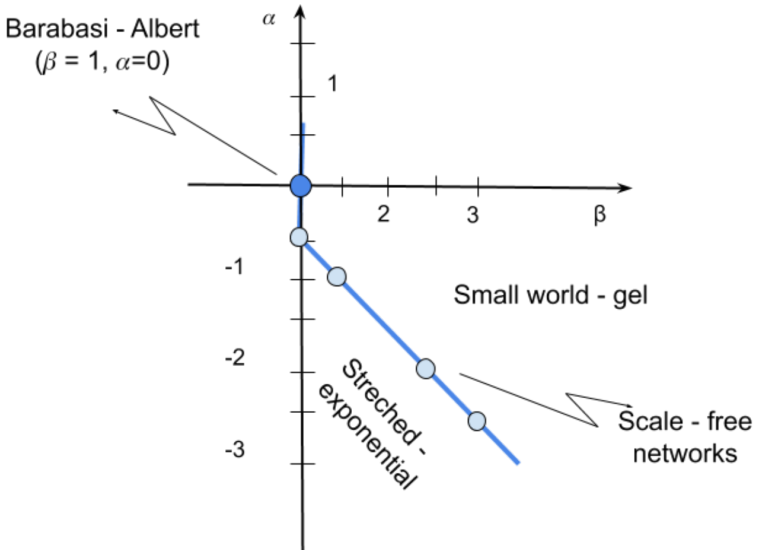

Figure 2.8: Phase diagram of aging network model.

2.5 Fractal analysis

The study of time series is an important approach in understanding complex systems [107], and the analysis of scaling laws and fractality in time series is particularly useful in characterizing their dynamics. With the Hurst exponent H, we can describe the degree of self-similarity or self-affinity across different scales of time in time series $x(t)$:

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

In other words, having self-similarity, means that if we rescale time t by a factor a, the time-series values $x(t)$ are rescaled by a factor a^H. Monofractal [108, 109] time series is characterized by a single scaling exponent that applies across all time scales. On the other hand, time series is called multifractal.

2.5.1 Long and short-term correlations

The autocovariance function $C(s)$ can be used to quantify the degree of persistence or correlation of a stationary time series [107], where the mean and variance do not change with time. The autocovariance function measures the linear dependence between the increments Δx_i and Δx_{i+s} at a lag s, where $\Delta x_i = x_i - x_{i-1}$, of time series $\{x_i\}, i = 1...N$, and it is defined as the expected value of their product:

$$C(s) = \langle \Delta x_i \Delta x_{i+s} \rangle = \frac{1}{N-s} \sum_{i=1}^{N-s} \Delta x_i \Delta x_{i+s}. \tag{2.41}$$

If the time series is uncorrelated, $C(s)$ is zero for all lags s. If the time series has short-range correlations, $C(s)$ decays exponentially with lag s, indicating that the correlations decay quickly with distance in time:

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

and this behavior is typical of time series generated by autoregressive processes,

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

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

If the time series has long-range correlations, $C(s)$ decays as a power-law with lag s, indicating that the correlations persist over long time scales. This behavior is typical of self-similar or fractal time series, and it is characterized by a power-law exponent γ such that:

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

Fourier filtering techniques can model this type of behavior. The Hurst exponent H is related to the power-law exponent γ by $H = 1 - \gamma/2$. Therefore, if we can estimate the Hurst exponent, we can infer the degree of persistence or long-range correlations of the time series.

Due to the presence of noise in the data and non-stationarity, directly calculating the autocovariance function $C(s)$ can be a challenging task. This is because non-stationarities make it difficult to define $C(s)$ properly, as its average may not be well-defined. Additionally, on large scales, $C(s)$ fluctuates around zero, which makes it impossible to determine the correct correlation exponent γ. Therefore, instead of computing $C(s)$, it is common to estimate the Hurst exponent H.

2.5.2 Rescaled range analysis

The rescaled range analysis (R/S) method proposed by Hurst [110]. is a popular technique to estimate the Hurst exponent of a time series. It is a simple method that works well for a wide range of self-similar processes. For time series x_i, we can define the profile Y_ν for each segment of the size s:

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

Constant trends in the data are removed by removing the average values over segment $\langle x_{\nu s+i}\rangle_s$. From there we can define the range between minimum and maximum value of obtained profile as $R_\nu(s) = maxY_\nu(j) - minY_\nu(j)$, and standard deviation is $S_\nu(s) = \sqrt{\frac{1}{s}\sum Y_\nu^2(j)}$.

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. The Hurst exponent can be estimated from the slope of the line in a log-log plot of $R(s)/S(s)$ versus s. Values $H < 1/2$ indicate long-term anti-correlated data while $H > 1/2$ long-term positively correlated data [107].

2.5.3 Fluctuation analysis

The fluctuation analysis is a method that relies on the principles of random walk theory [107]. It involves taking a time series x_i of length N and creating a global profile by calculating the cumulative sum using equation 2.42. In this equation, $\langle x\rangle$ represents the average value of the time series.

$$Y(j) = \sum_{i=0}^{j}(x_i - \langle x\rangle), \quad j = 1, ..., N. \tag{2.42}$$

Figure 2.9 shows examples of multifractal, monofractal and white noise signal with their global profiles.

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

$$F_2(s) = [\frac{1}{2N_s}\sum F^2(\nu, s)]^{1/2} \sim s^H. \tag{2.43}$$

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.

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

Figure 2.10 shows the global profile of the multifractal signal, divided in segments of the length $s = 1000$. On the top panel, each segment s is approximated with linear function.

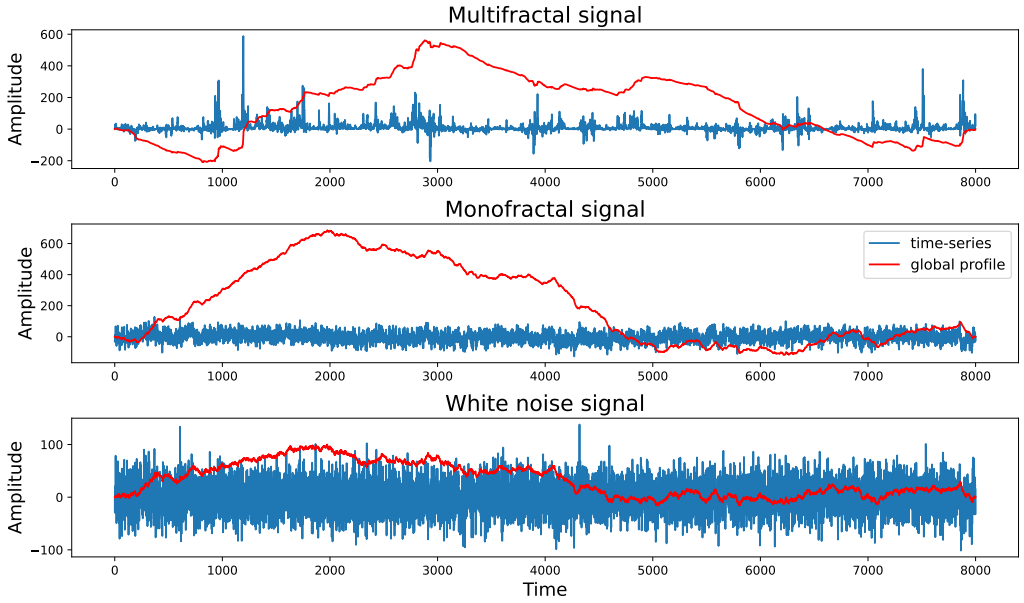

Figure 2.9: Multifractal, monofractal and white noise signals.

The trends present in the time series do not have to be linear [111]. The middle and bottom panel in Figure 2.10 show that the segments of the signal could be very well approximated with some higher order functions: quadratic or cubic. In general, using the detrended fluctuation analysis (DFA) we could remove the polynomial trend of the order m [112]. 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:

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \left[Y(j) - p^m_{\nu,s}(j) \right]^2 . \tag{2.44}$$

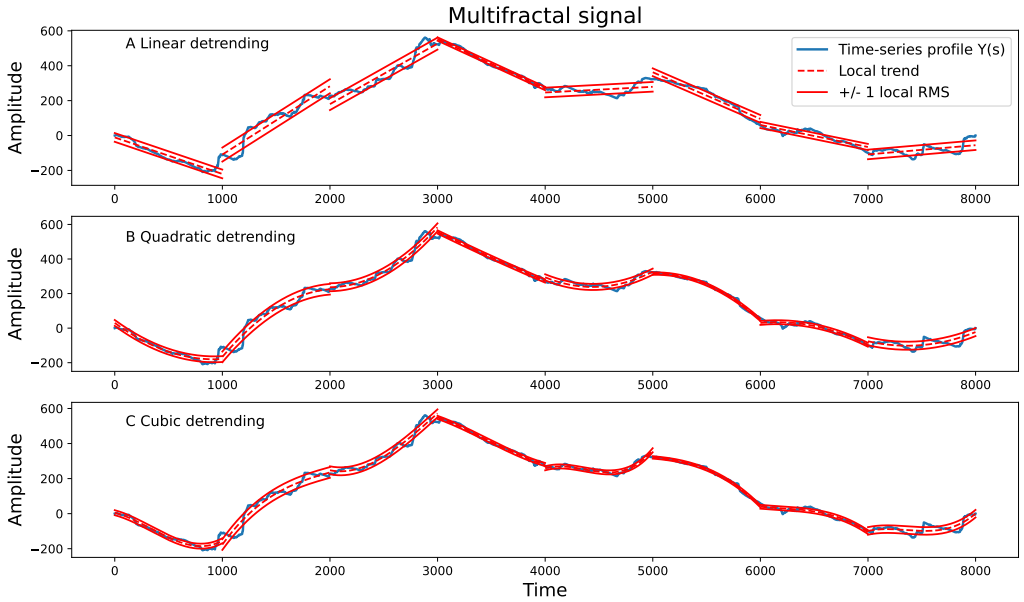

Figure 2.10: Detrending of multifractal signal for the segments of length $s = 1000$. Panel A- linear detrending, panel B-quadratic detrending, panel C- cubic detrending.

2.5.4 Multifractality of the signals

The scaling behavior in many data may be more complicated, resulting that interwoven subset of time series have different scaling exponents. This property is known as multifractality. The multifractality may be caused by the time series values' large probability distribution [113, 114]. In this situation, shuffling time series cannot eliminate the multifractal features. The source of multifractality may also come from different small and large fluctuations correlations. If density function is distribution with finite moments, the shuffled time series will lose multifractal properties as correlations are easily destroyed with randomization. In situations where multifractality is caused by both types, the randomized time series has weaker multifractality.

Multifractal detrended fluctuation analysis (MFDFA) is used [113, 114] 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 $F(q), q- > 0$, cannot be calculated directly because the exponent diverges. Instead, the logarithmic averaging procedure has to be considered.

$$F_0(s) = \exp \left\{ \frac{1}{4N_s} \sum_{\nu}^{2N_s} ln \left[F^2(\nu, s) \right] \right\}, q = 0. \tag{2.45}$$

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 2.11.

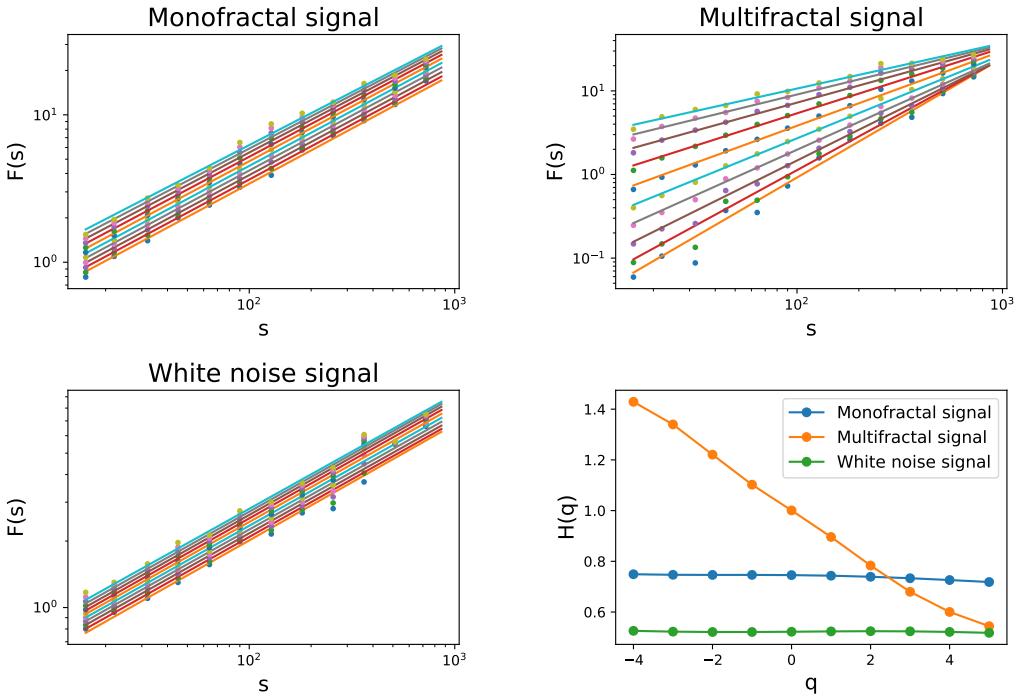

Figure 2.11: Dependence of the fluctuating functions on the scale for monofractal, multifractal and white noise signals, and the Dependence of the Hurst exponent H on the scale 1 q for different types of signal (bottom right).

For monofractal time series, the scaling properties of all segments are the same, regardless of their size or magnitude of change. This means that the value of H(q) will be the same for all values of q [113, 107]. If the series exhibits multifractal behavior, then the scaling properties of different segments of the series will be different, and the value of H(q) will vary depending on the magnitude of change in the segment being analyzed. Positive values of q will indicate segments with large fluctuations, while negative values of q will describe the scaling of segments with small fluctuations [107].

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' behavior 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 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) [67], we can quantify the user reputation R_n after each interaction using equation 2.46, where n is the number of interaction $n \in 1, N$

$$R_n = R_{n-1}\beta^{\Delta_n} + I_n. \tag{2.46}$$

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}:

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a}, \tag{2.47}$$

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

$$I_n = I_{b_n}(1 + \alpha(1 - \frac{1}{A_n + 1})). \tag{2.48}$$

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 2.12 - 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 behavior, 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,

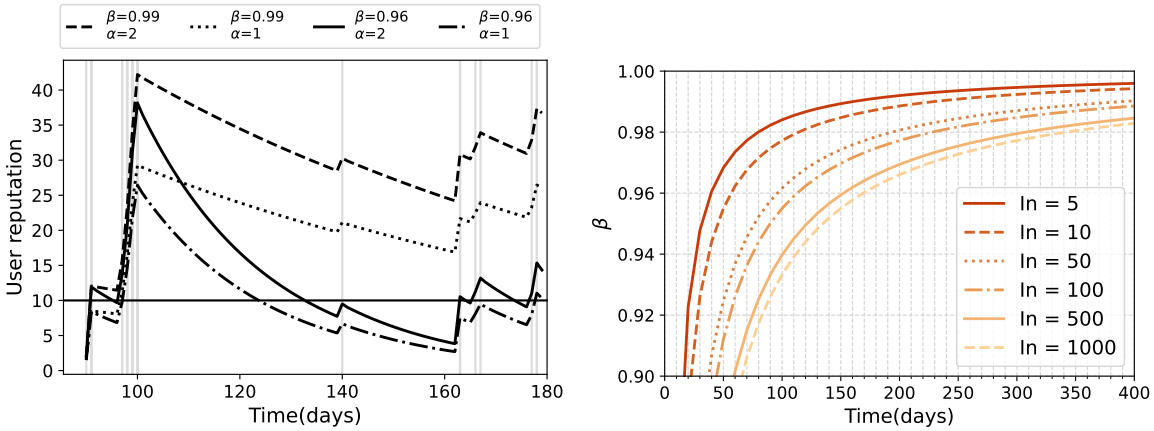

Figure 2.12: Left panel shows the dynamics of user reputation obtained in DIBRM model for different model parameters α and β. Right panel shows the dependence of parameter β and number of days to reputation from starting value I_n drops below threshold $I_n = 1$.

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

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 2.12, - 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 reputation has highest growth when user shows burst in activity. 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. In this thesis DIBRM model is used to analyze Stack Exchange communities, Chapter 5, while in Appendix B, we suggest the procedure to estimate the model parameters for this specific system.

Chapter 3

Evolving complex network structure dependence on the properties of growth 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 power-law degree distribution in the Barabasi-Albert model [32]. 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 [115, 116]. The dynamics 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 [108].

In this chapter, we study how growth signals influence the network structure. We explain the properties of growth signals, both real and computer-generated, and analyze networks created with a growing network model where the interplay between aging 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:

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

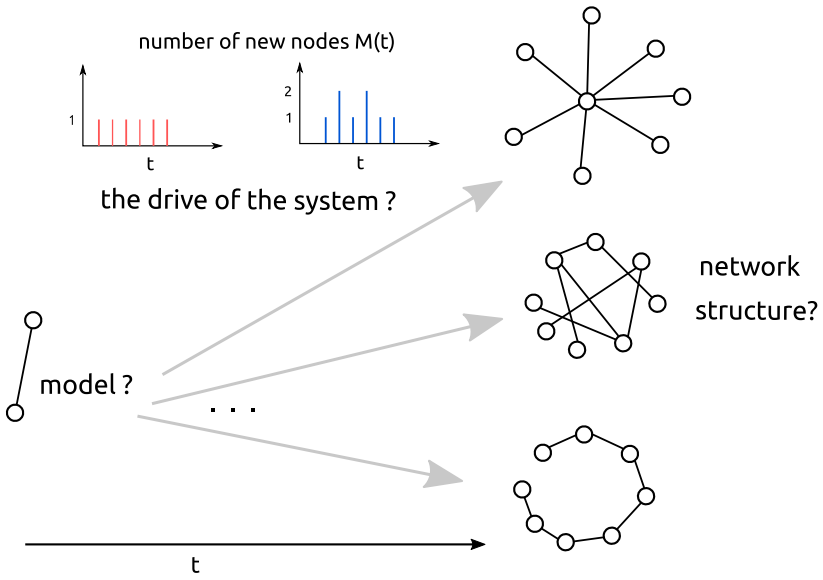

Figure 3.1: The open question is how nonlinear signals, in combination with the network model, influence the network's structure. Under what circumstances do networks have the scale-free, hub-spoke, or chain structure?

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, 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 [34]. 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** [117] community in the Meetup website and on two months of **MySpace** [118] social network. Besides these signals, we use randomized MySpace and TECH signals and uncorrelated Poissonian signals.

3.1.1 Characteristics of growth signals

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 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.

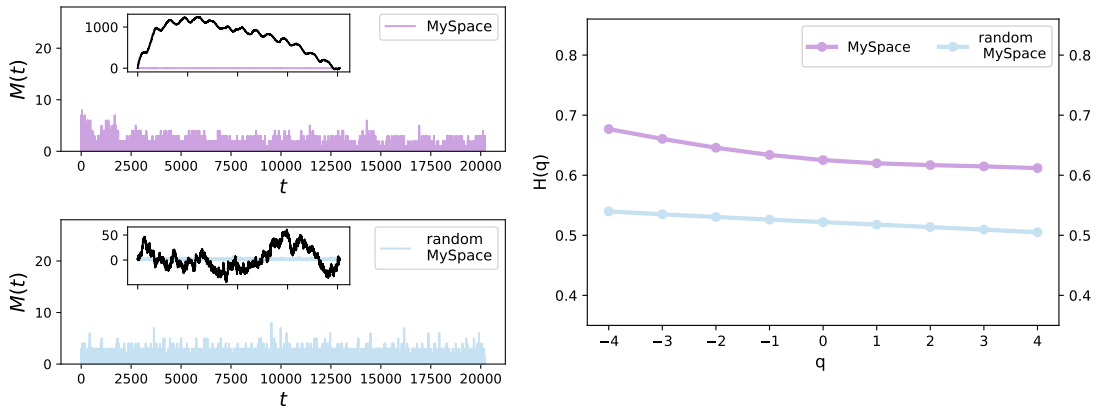

Figure 3.2: MySpace signal, the random MySpace signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q (right pane).

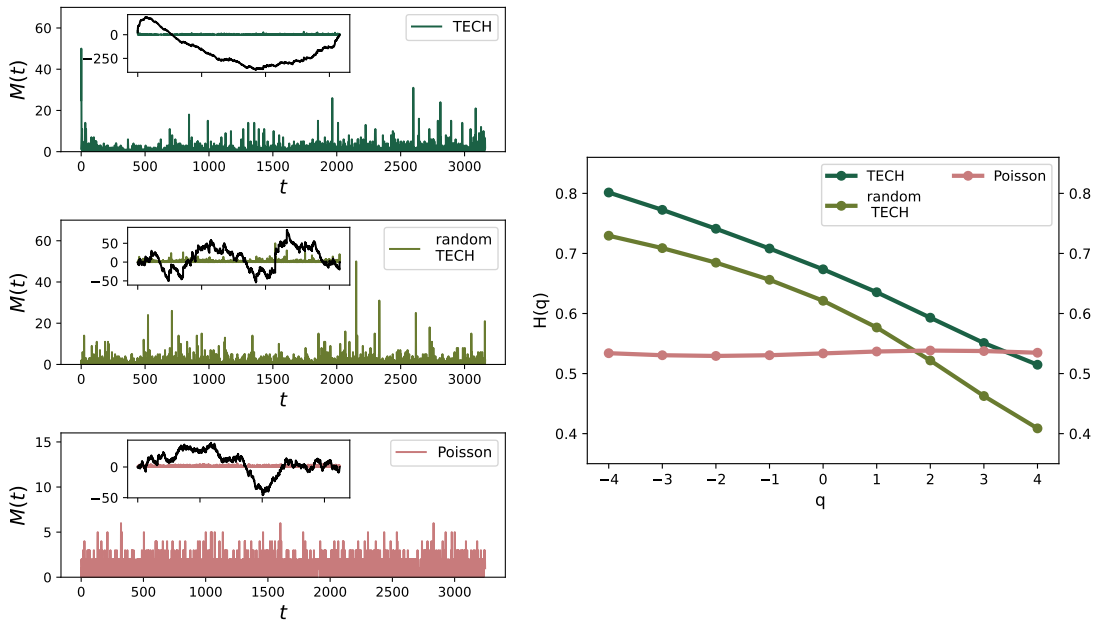

Figure 3.3: TECH signal, the random TECH signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q (right pane).

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 Structural differences between evolving complex networks

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 [76], 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 [76].

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 behavior 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 the 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 $L = 2$, $\alpha = -1.0$, $\beta =$

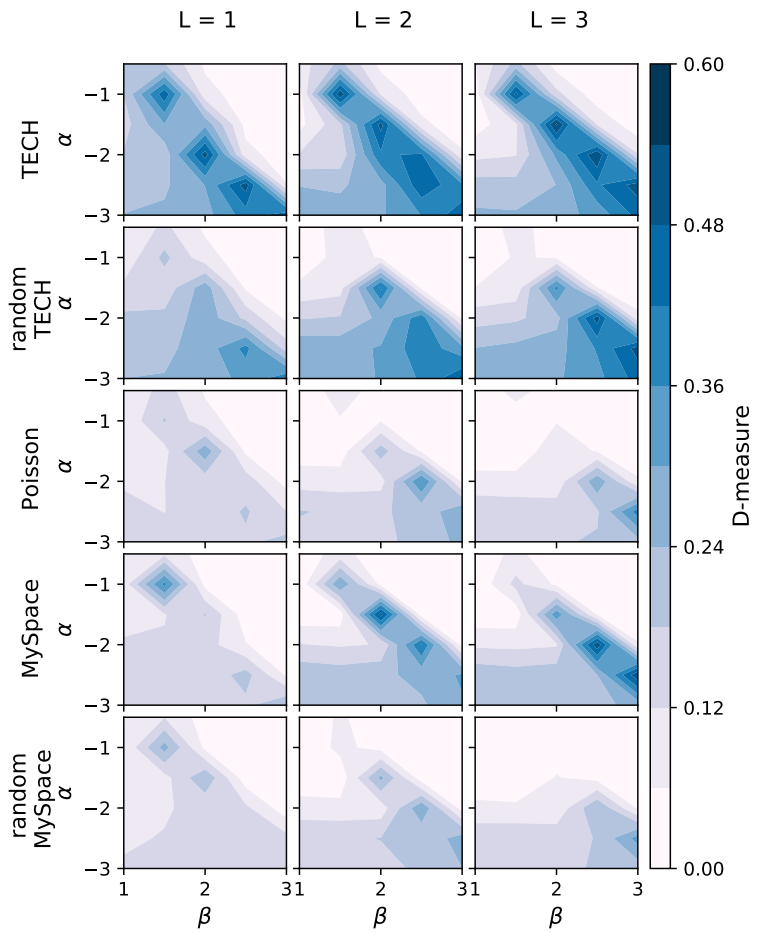

Figure 3.4: The comparison of networks grown with growth signals shown in figures 3.3 and 3.2 versus ones grown with constant signal $M = 1$, for the value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in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

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 neighboring 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 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.

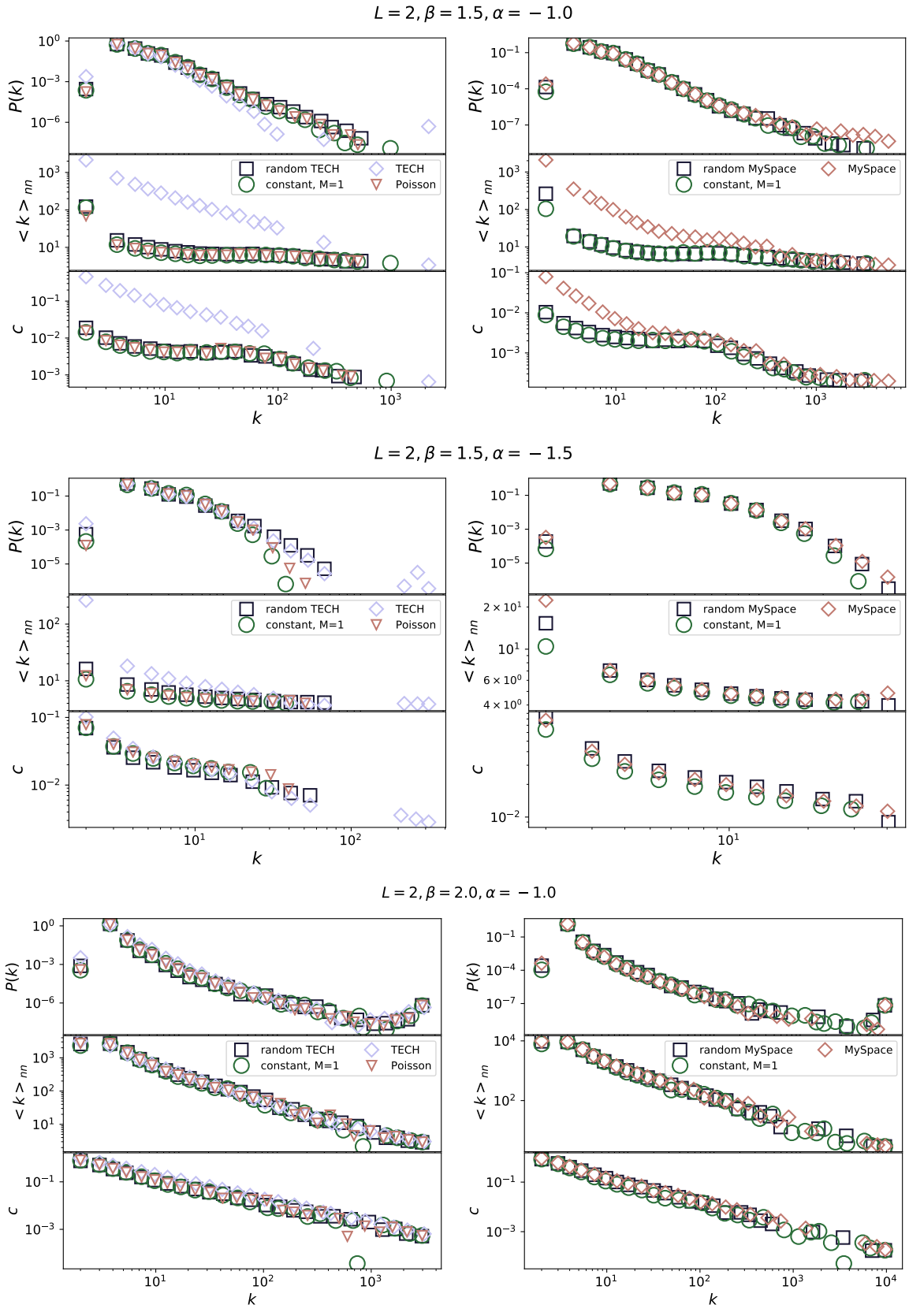

Figure 3.5: Degree distribution, the dependence of average first neighbor degree on node degree, the dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $\alpha = -1.0$, $\beta = 1.5$ and $L = 2$ for all networks. 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 [119].

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 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.

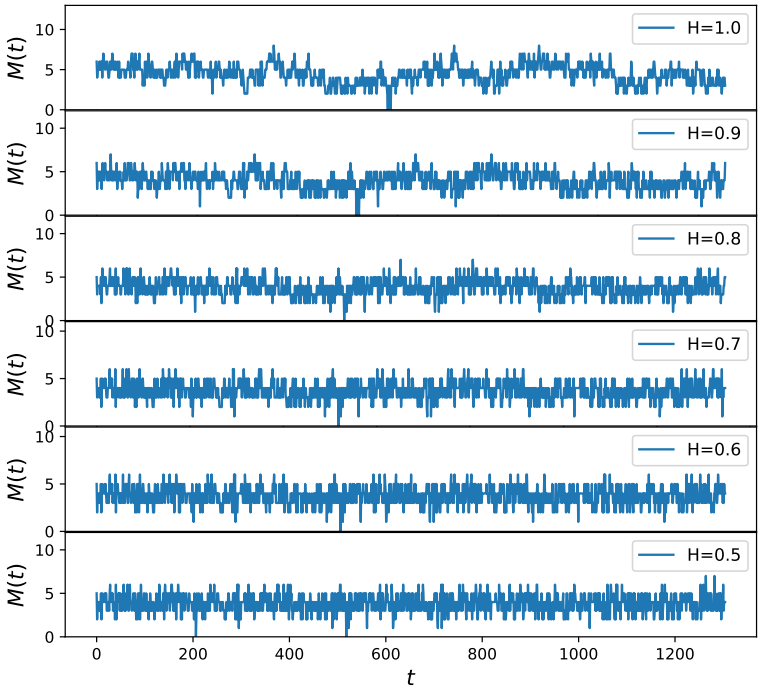

Figure 3.6: Monofractal signals generated with Fourier filtering method for different Hurst exponents

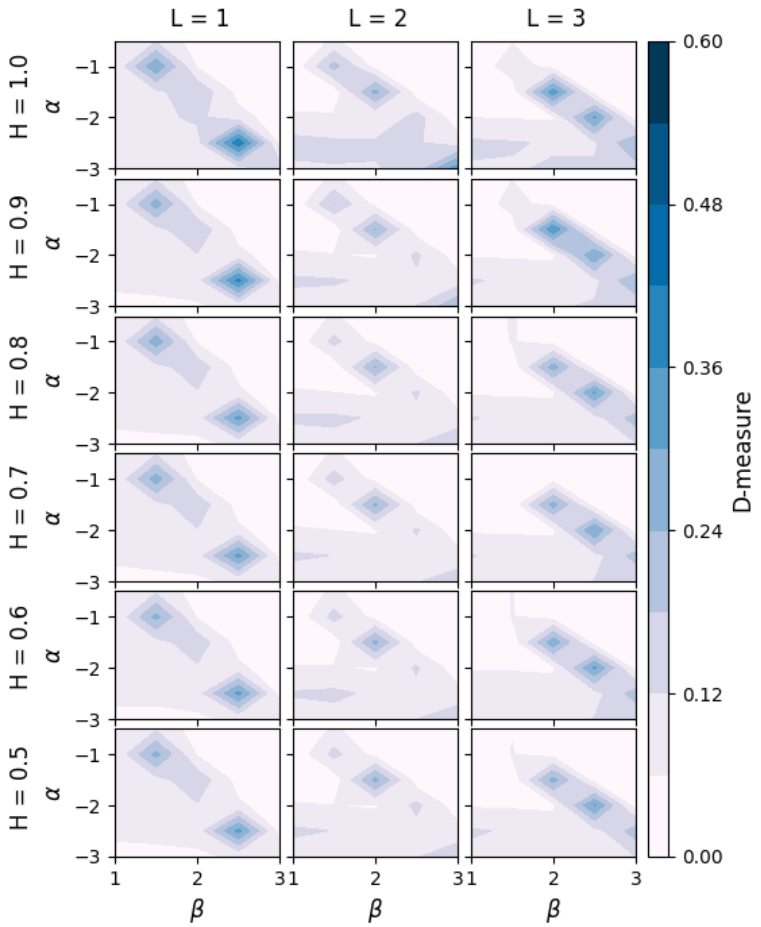

Figure 3.7: D-distance between networks generated with different long-range correlated signals with a fixed value of Hurst exponent and networks generated with constant signal M=4.

We further explore the assortativity index and clustering coefficient of generated networks. Figure 3.8 are results for several aging 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.

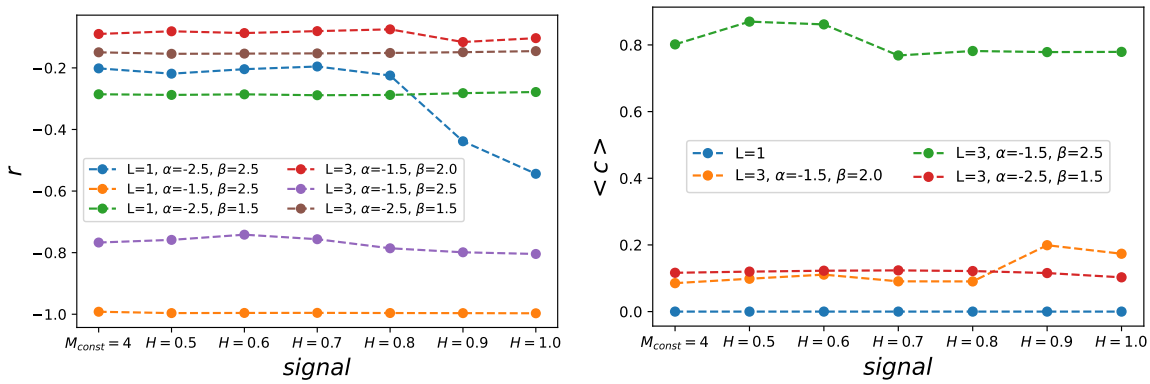

Figure 3.8: Mean assortativity index for networks generated with different model parameters α, β, L and different long-range correlated signals with Hurst exponent H.

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 the 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 signals and signals 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 aging 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 aging 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 have a strong influence on 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 [54] has shown that edge activation properties impact the complex system's dynamics. Also, different studies indicated the importance of fluctuating signals [27, 40, 35]. Our results imply that modeling 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

The evolving complex networks have a tendency to separate into connected fragments, communities, or groups of nodes. These communities are formed around certain topics and interests; they could also evolve and influence network structure and members' behavior. The distribution of the sizes of these communities has a universal shape. To understand how the dynamics and structure of the networks affect the distribution of community sizes, we combine empirical approaches and theoretical modeling. We analyze real-world social networks and collect data about their structure and community sizes, while theoretical modeling involves developing models able to capture essential features of social networks and explain the emergence of the universal distribution of group sizes.

4.1 Empirical analysis of the social group growth

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

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

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 the group. So both datasets have the same structure: (g, u, t), where t is the timestamp when user u 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 $R = log(S_i(t)/S_i(t-1))$.

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 [99]. 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, [120], [121]. 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 [122, 123].

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.

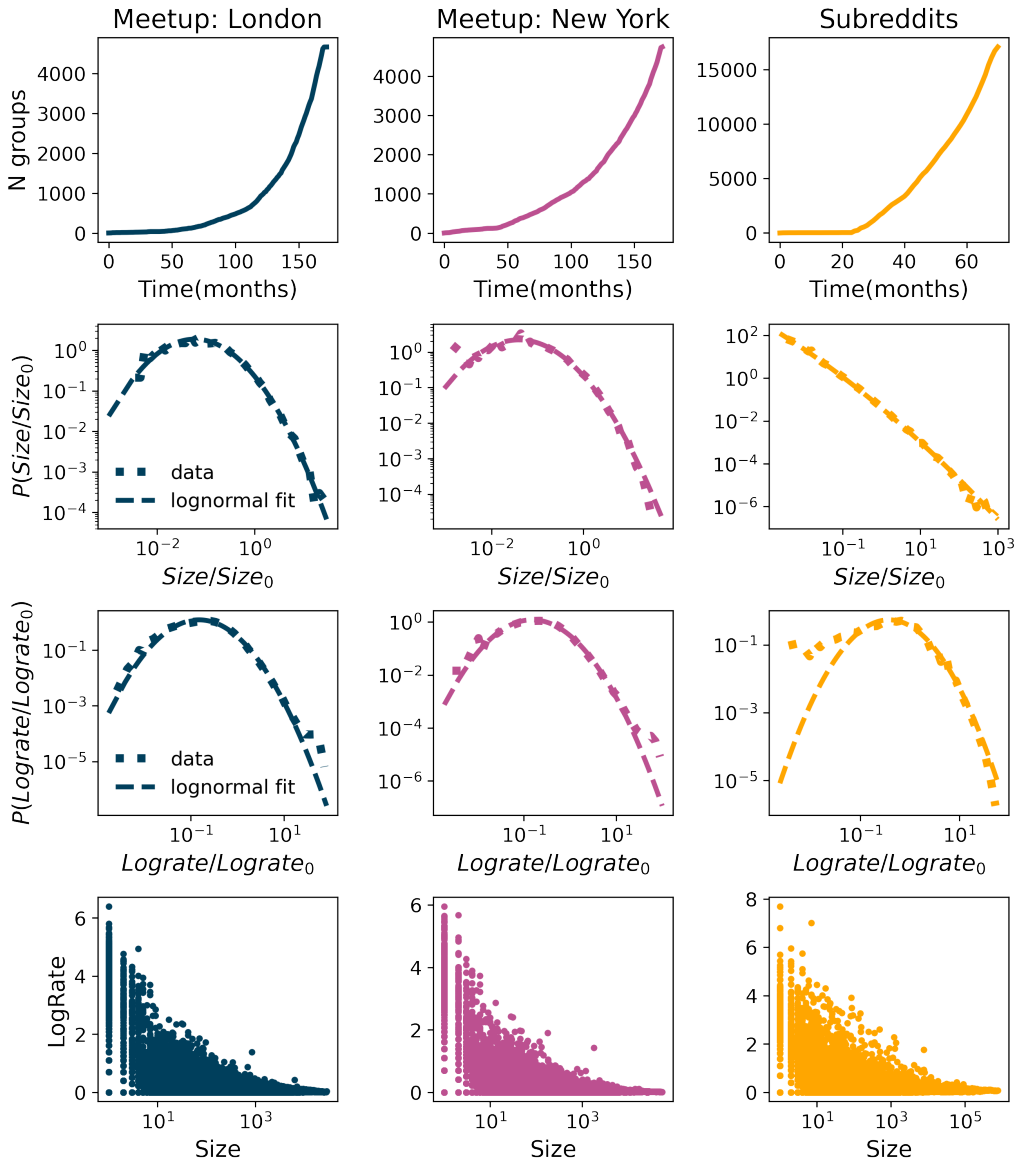

Figure 4.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011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

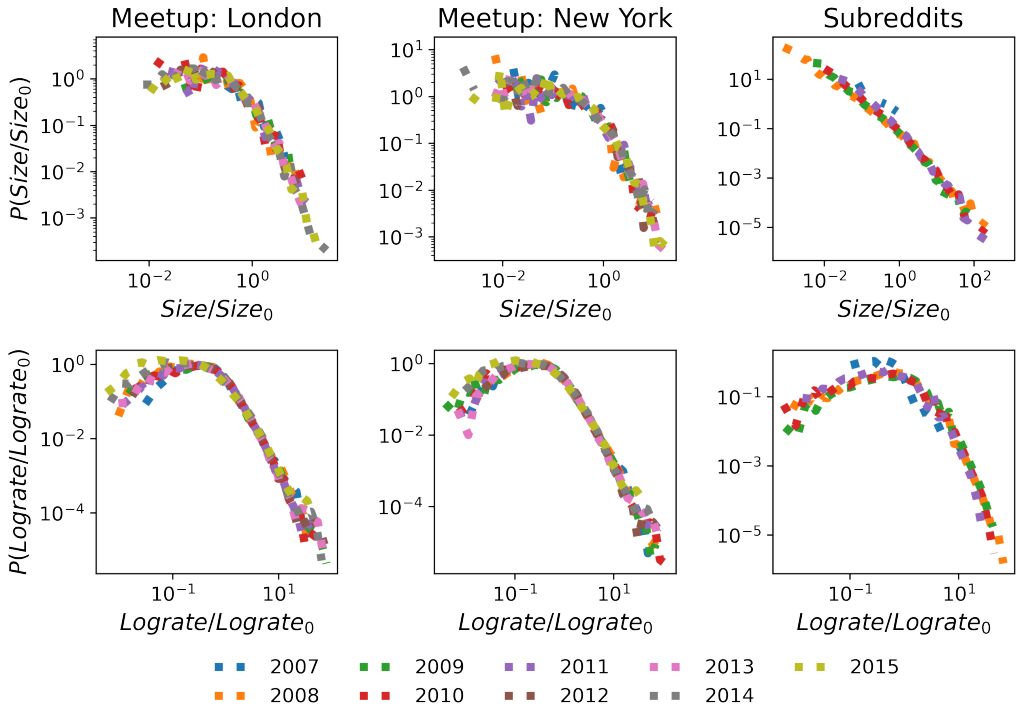

Figure 4.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

4.2 Theoretical model of social group growth

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 modeled 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 [124, 125]. 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 [126, 127, 63].

The evolution of the social groups has been studied using the co-evolution model in the reference [125]. 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 [124]. 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.

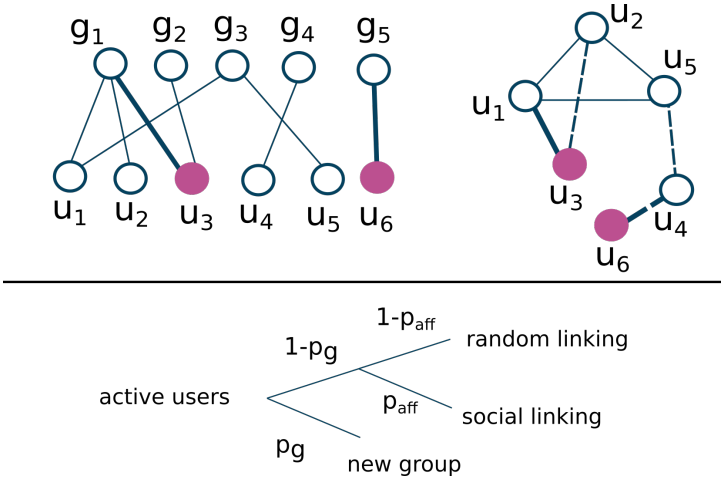

Figure 4.3: The top panel shows bipartite (member-group) and social (member-member) networks. Filled nodes are active members, while thick lines are new links in this time step. In the social network, dashed lines show that members are friends but do not share the same groups. The lower panel shows the 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 a group g_1, member u_3 will also join group g_1. When member u_3 joins 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 neighborhood 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 neighborhood 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 [117, 128, 125] 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.

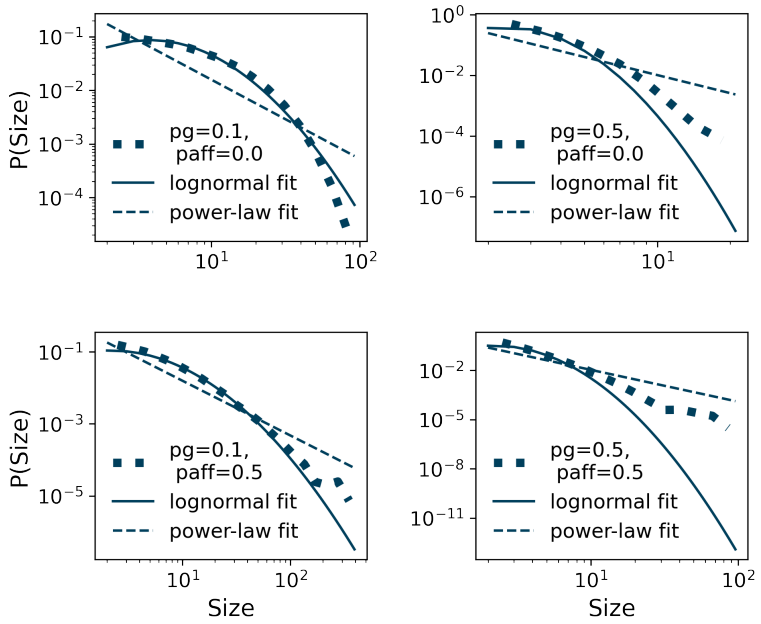

Figure 4.4: The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

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 [125], 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.

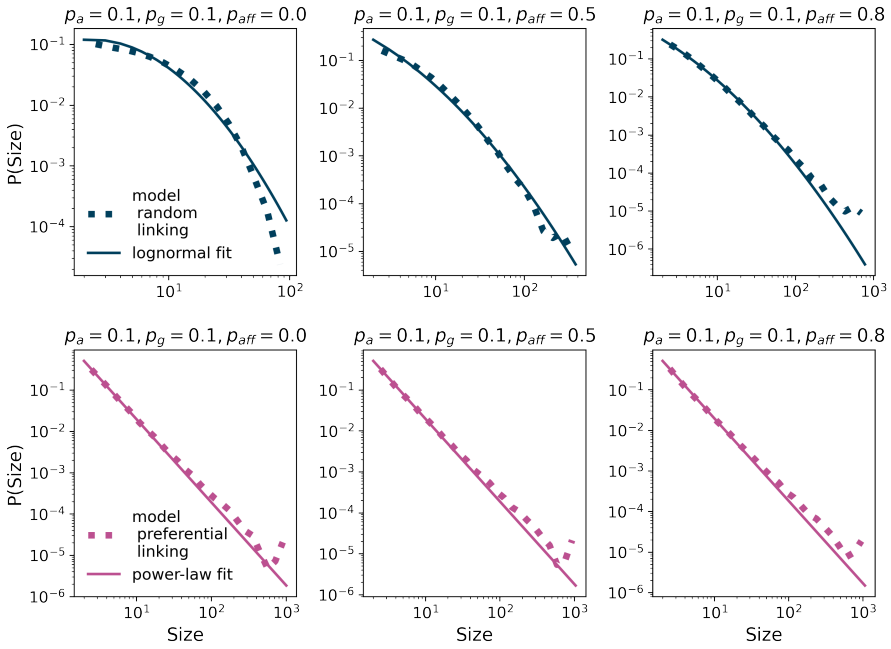

Figure 4.5: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff}. With probability, $1 - p_{aff}$, 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 The growth of real social groups

The social systems do not grow at a constant rate. In Ref. [63], 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.

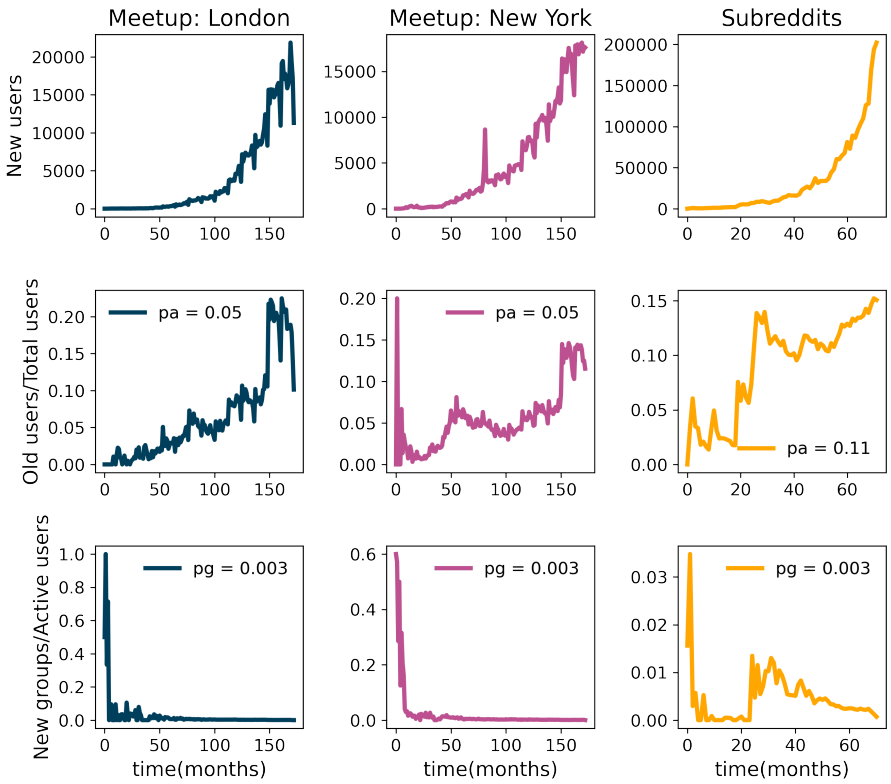

Figure 4.6: The time series of the number of new members (top panel), the ratio between old members and total members in the system (middle panel), and the ratio between new groups and active members(bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

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 [129] 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.

Table 4.1: Jensen Shannon divergence between group sizes distributions from model (in the model, we vary affiliation parameter paff) and data.

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

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 that are the most likely consequence of using median values of parameters p_a, p_g, and p_{aff}.

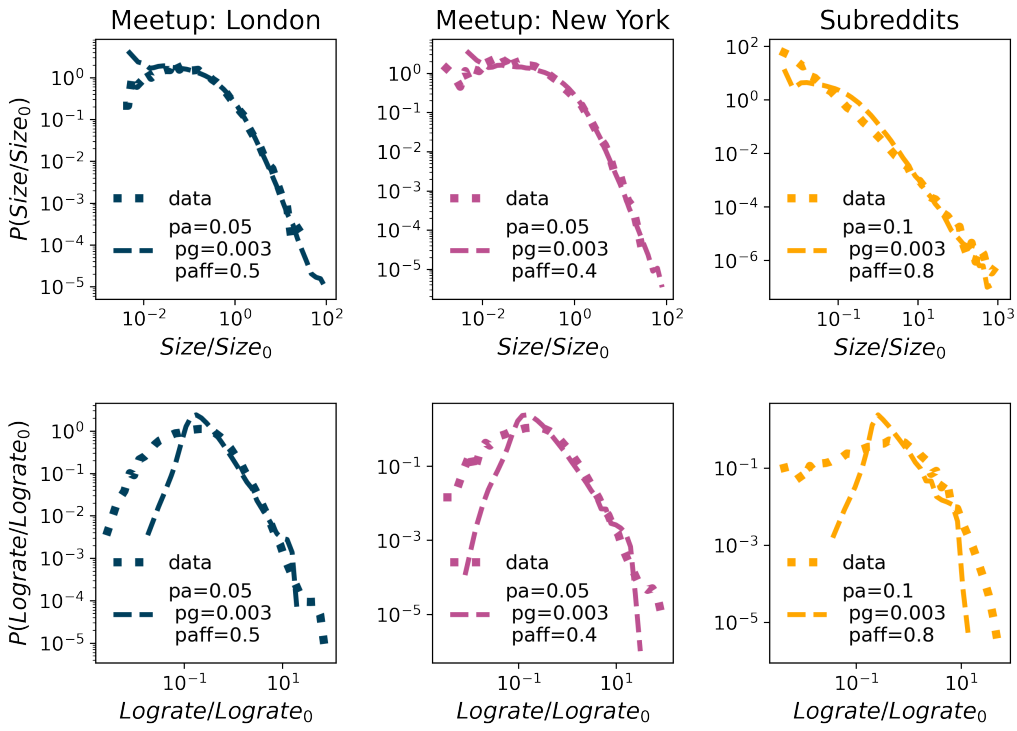

Figure 4.7: The comparison between empirical and simulation distribution for group sizes (top panel) and lograte (bottom panel).

4.3.1 Distributions fit

We compute the log-likelihood ratio R and p-value between different distributions and lognormal fit [103] 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 [103]. 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

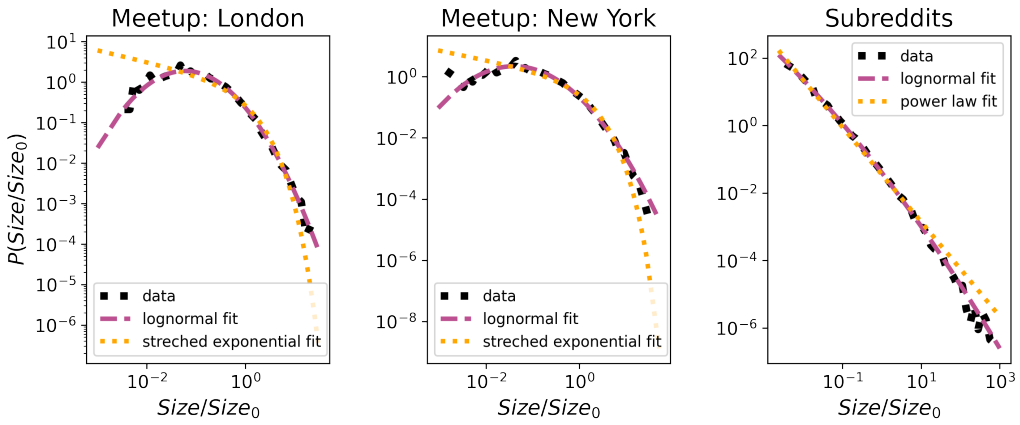

Figure 4.8: The comparison between lognormal and stretched exponential fit to London and NY data, and between lognormal and power law for Subreddits. The parameters for lognormal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits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

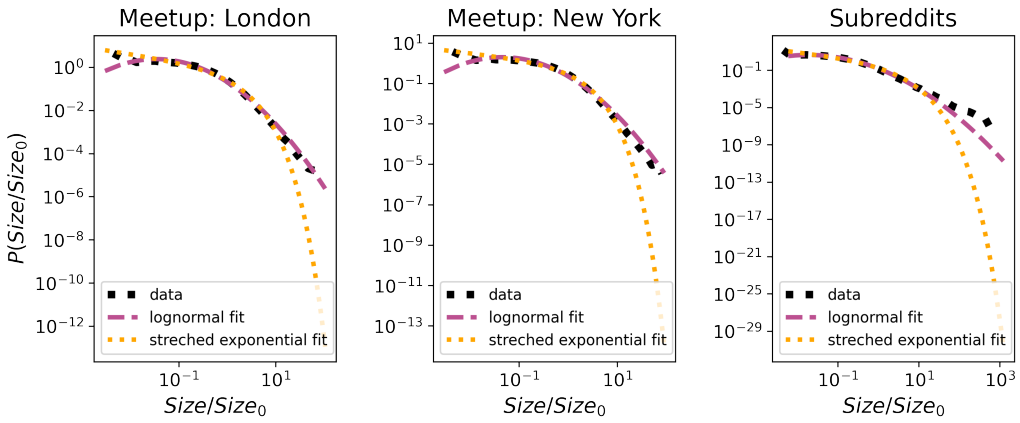

Figure 4.9: The comparison between lognormal and stretched exponential fit to simulated group size distributions. The parameters for lognormal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma =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

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 [130], and the groups' growth model produces the same property. In Figure 4.10, blue dots show degree distributions of modeled 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 [131], 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.

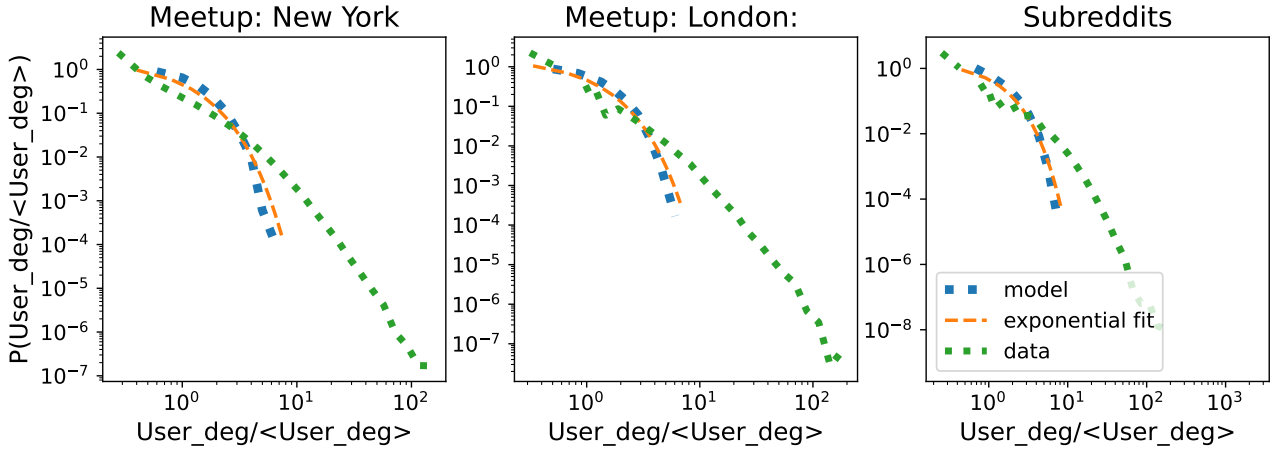

Figure 4.10: Users degree distributions from empirical data are compared to degree distributions observed by groups growth model.

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 behavior. Contrary, the previous work on online social groups indicated that group size distributions of LiveJournal and Youtube follow power-law [125]. 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 used a model where the social system is presented with evolving bipartite and social networks [125]. 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 also decide to be 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 estimated model parameters p_a, p_g, and p_{aff} from empirical data. We saw 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 [125]. 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 modeled.

Chapter 5

The sustainability of evolving knowledge-based communities

One of the key findings from the research on complex networks is that the structure of social interactions plays a significant role in their sustainability [117, 132]. Social interactions can be positive and negative, playing a vital role in shaping network dynamics. Positive interactions, such as cooperation, can lead to the formation of clusters or communities within the network, promoting its sustainability [133, 134]. In contrast, negative interactions, such as competition, can lead to the breakdown of the network structure and decrease its sustainability [135, 132]. Social interactions can also influence the emergence of collective behavior, which can significantly impact its sustainability [117, 132]. In this chapter, we study Stack Exchange communities' structure, dynamics, and sustainability.

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 Stack Overflow [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 the 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.

[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.

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

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 [12], and it was shown that degree distribution, degree-degree correlations, and clustering coefficient are sufficient for the description of the properties of complex networks [136].

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.

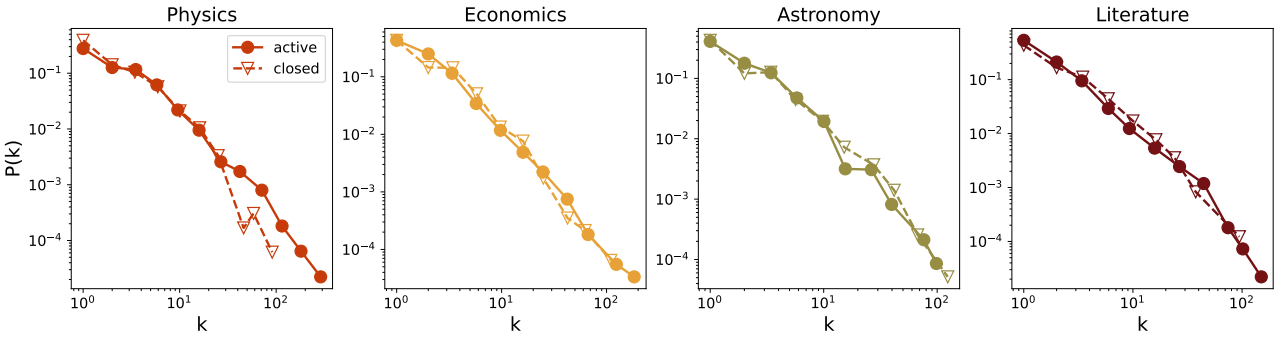

Figure 5.1: Degree distribution of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 neighbors 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 behavior, where the $k_{nn}(k)$ distributions of closed communities are on top of the distributions of the active ones.

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 [128].

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

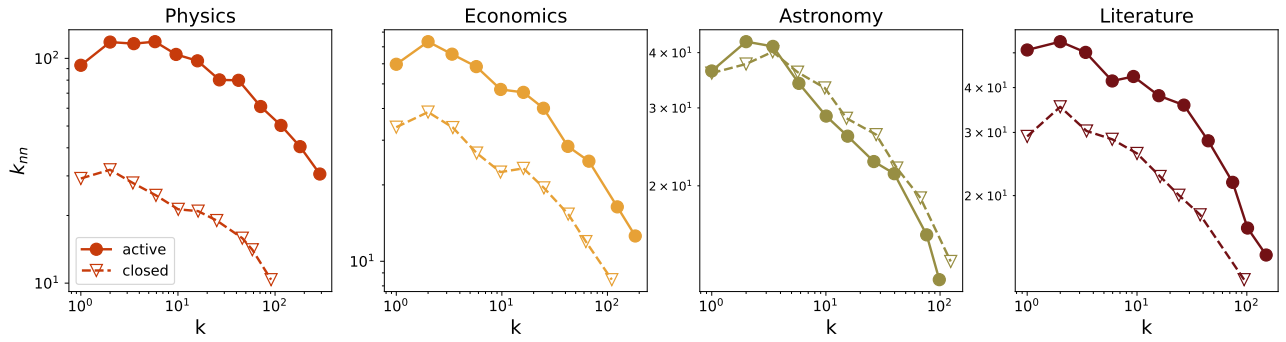

Figure 5.2: Neighbor degree dependence on the node degree of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

Furthermore, successful social diffusion typically occurs in networks with a high value of clustering coefficient [137]. 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.

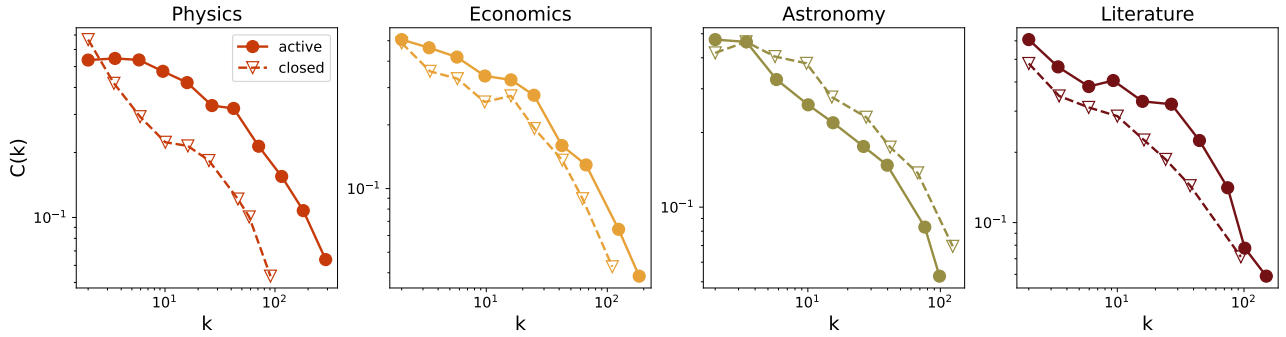

Figure 5.3: Clustering coefficient dependence on the node degree of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 behavior 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

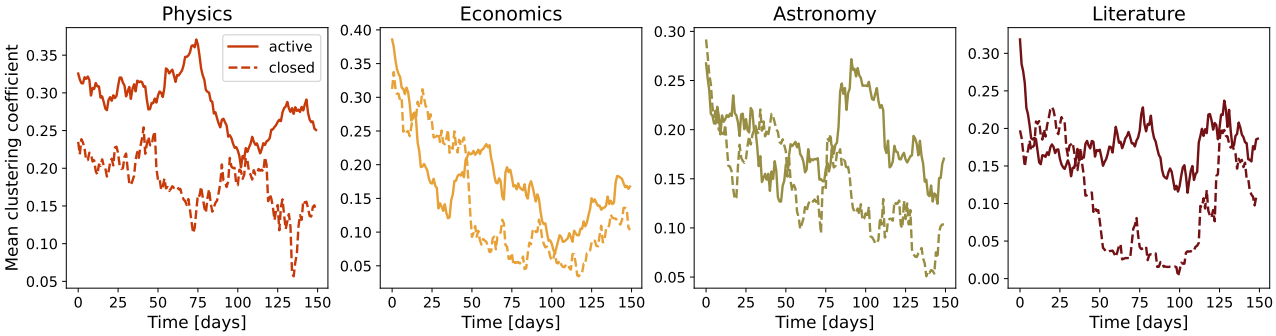

Figure 5.4: Mean clustering coefficient of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

phase, while later communities that are still active have higher clustering coefficient values. These 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 [138, 139]. 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.

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 [77, 66].

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, [66]. 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

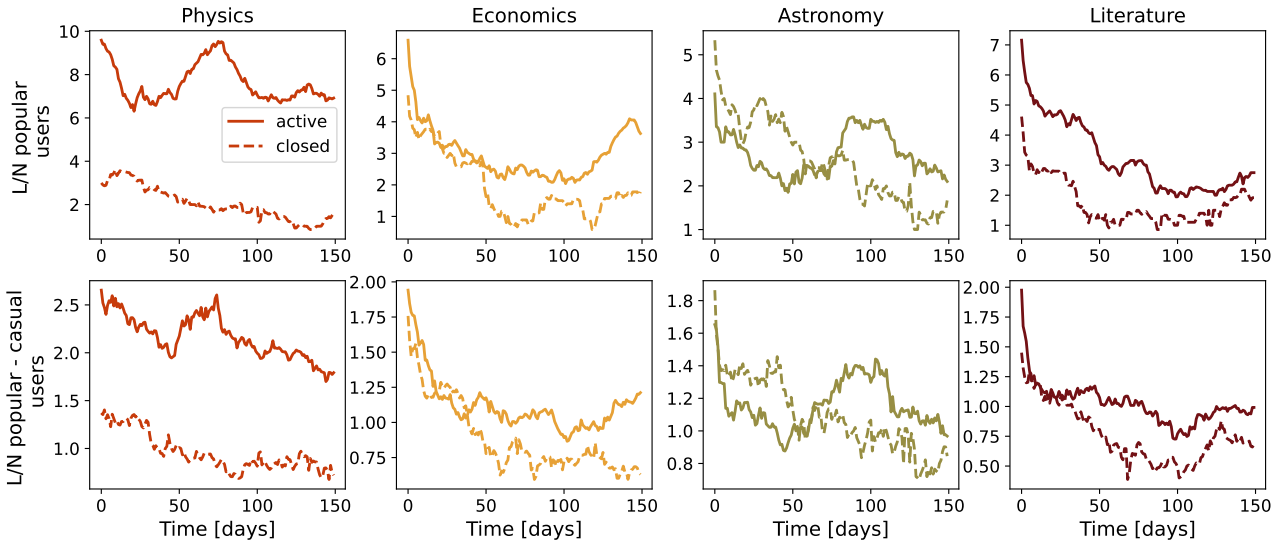

Figure 5.5: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users) of four pairs four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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.

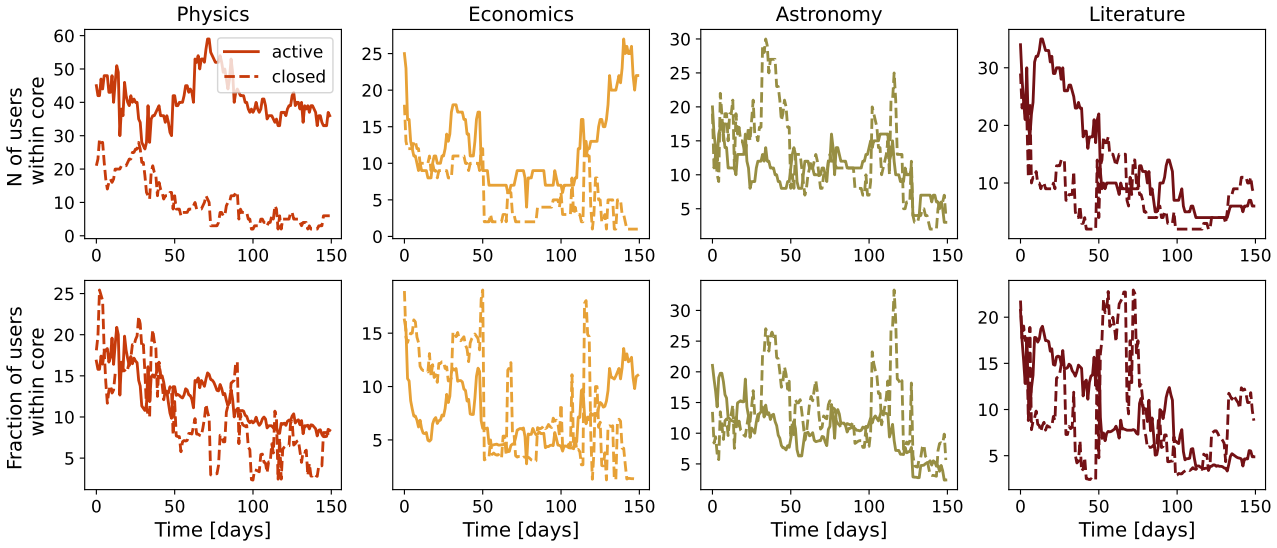

Figure 5.6: The size of the core (top) and a fraction of users in the core (bottom) of four pairs four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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.

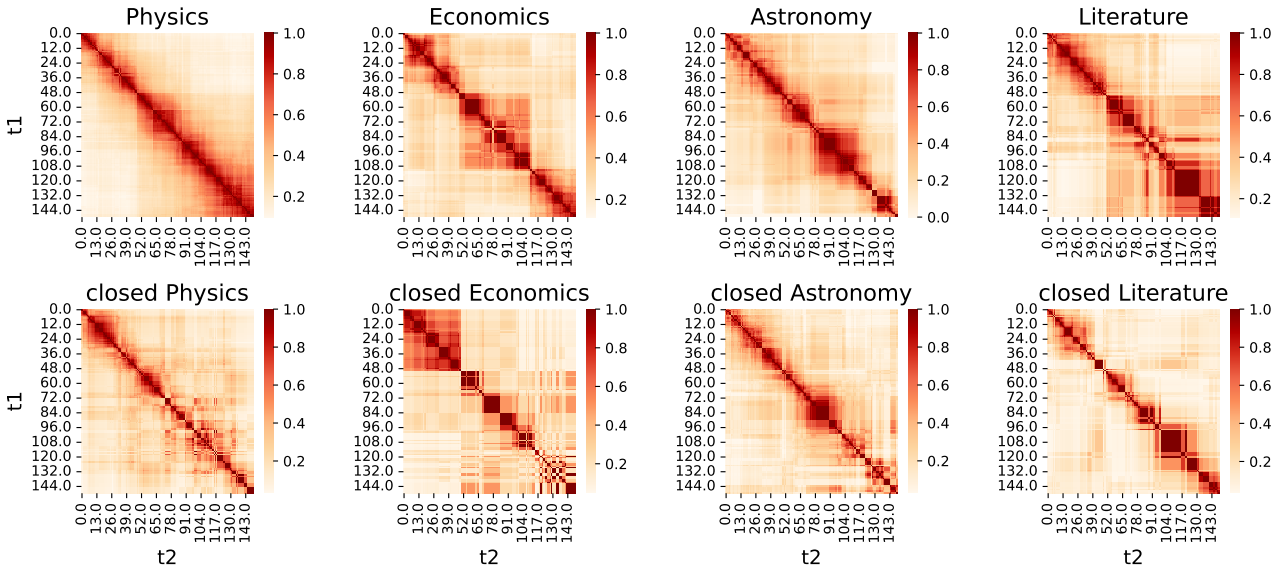

Figure 5.7: Jaccard index between core users in sub-networks at time points $t1$ and $t2$ for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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.

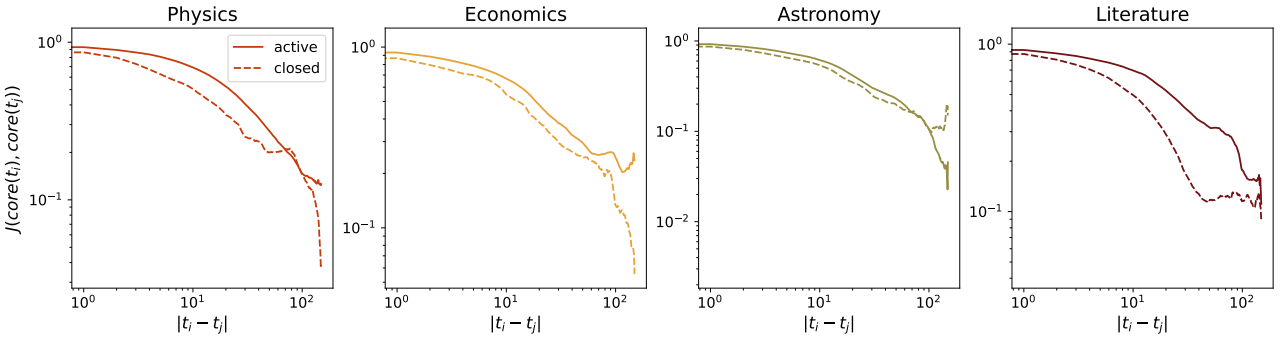

Figure 5.8: Jaccard index between core users in 30 days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$ for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

Finally, we examine how the users' connectivity in and between the core and periphery evolves. In 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.

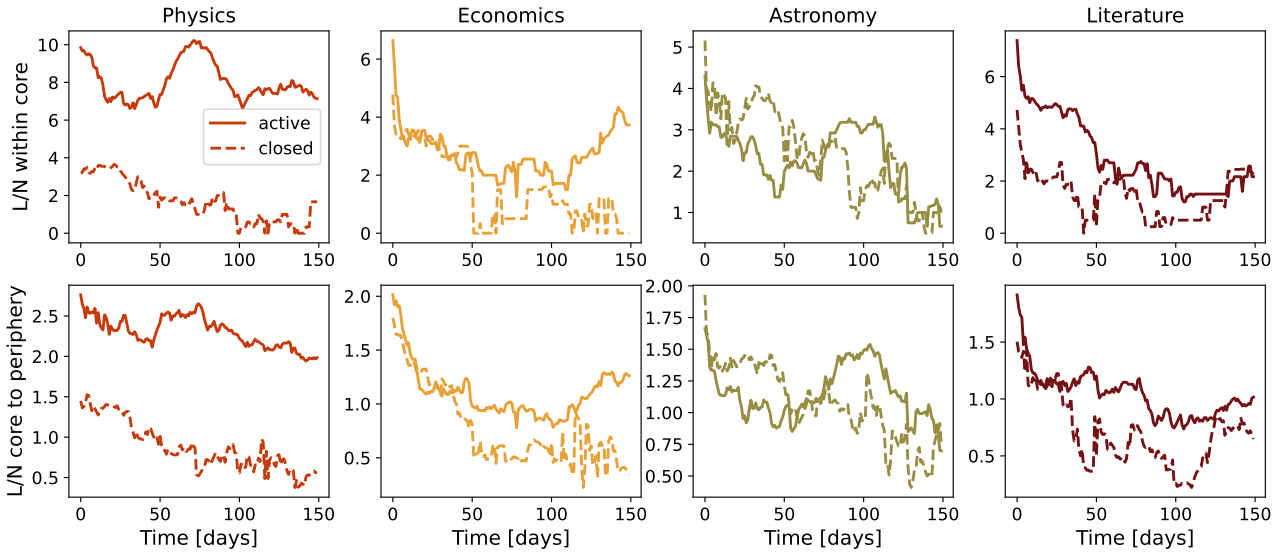

Figure 5.9: Number of links per node in core (top panel) and between core and periphery (bottom panel) for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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.

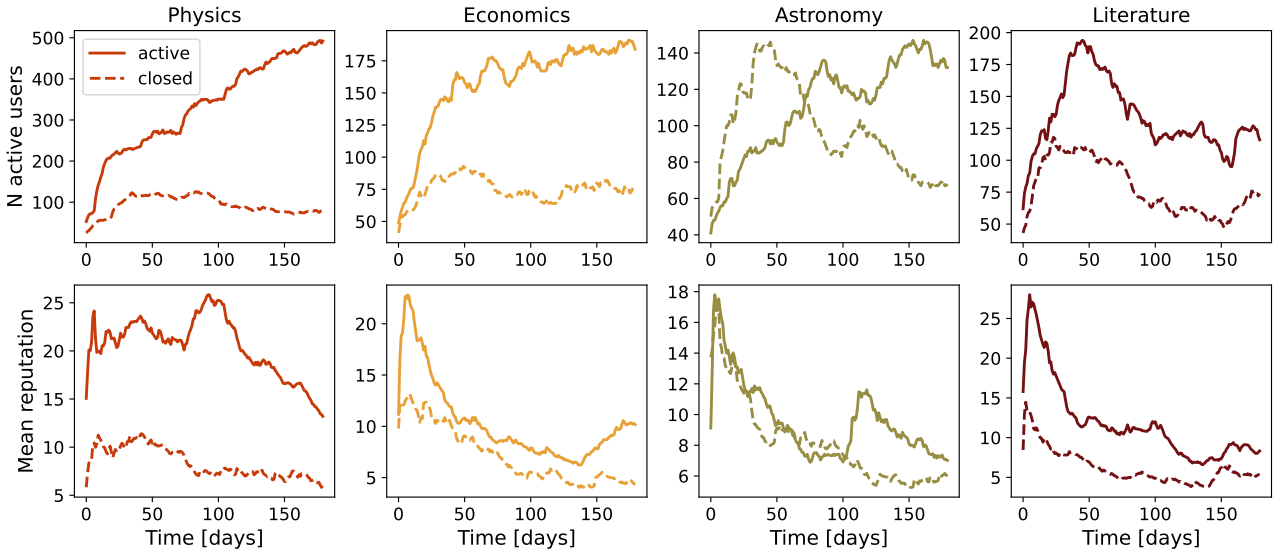

Figure 5.10: Number of active users (top panel) and mean reputation (bottom panel) of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 B.

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.

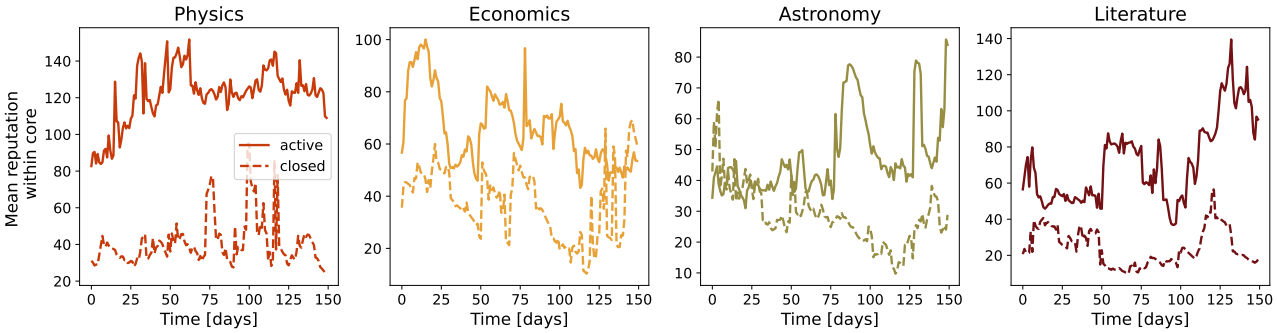

Figure 5.11: Dynamical reputation within the core of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 favor of active Economics. Astronomy shows different behavior where the closed community where dominant; closed astronomy had a larger core-periphery reputation ratio.

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.

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

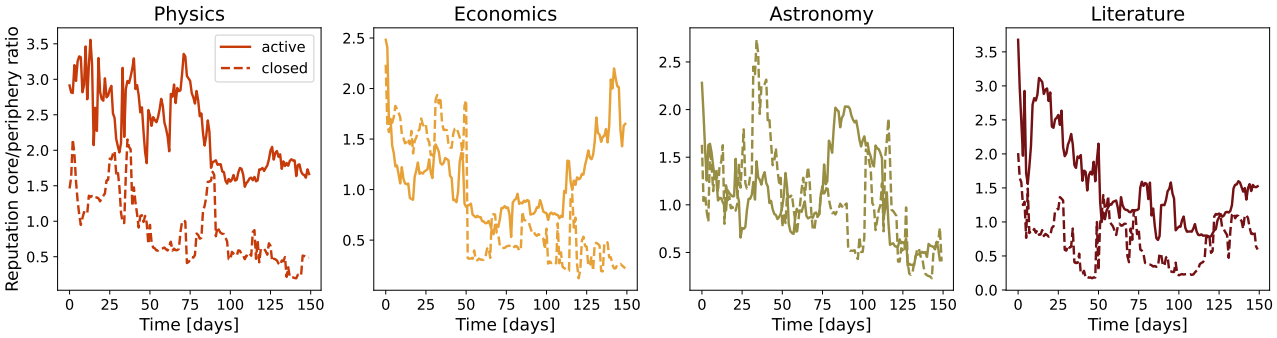

Figure 5.12: Ratio between the total reputation within network core and periphery of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

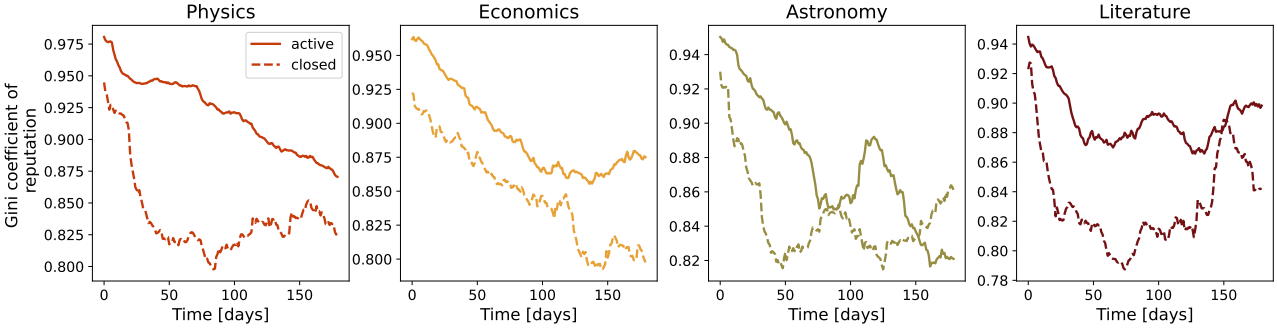

Figure 5.13: Gini index of dynamic reputation within the population of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 behavior: 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.

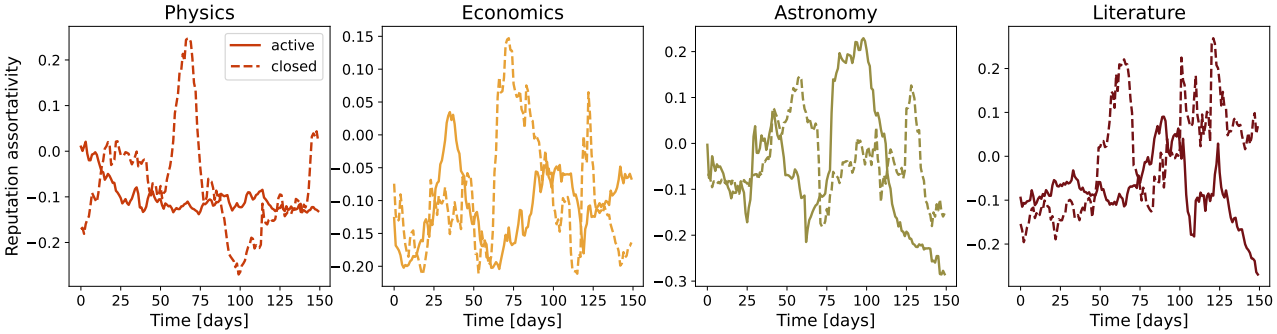

Figure 5.14: Dynamic Reputation assortativity in the network of interactions of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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.

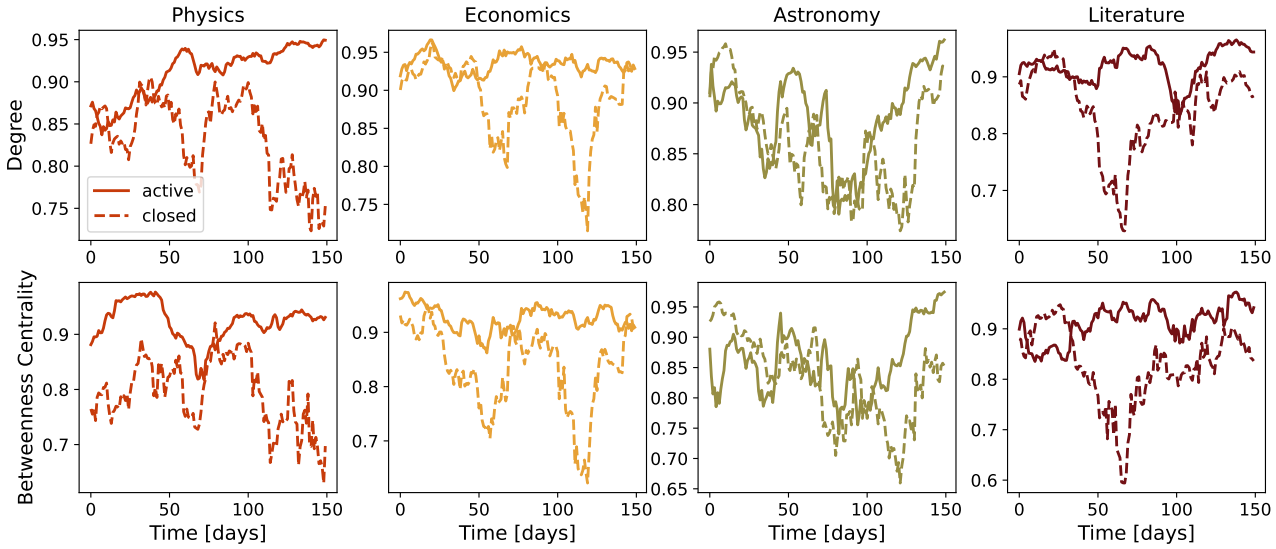

Figure 5.15: Coefficient of correlation between users' Dynamic Reputation and users' network degree (top) and users' betweenness centrality (bottom) of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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 fulfill 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 commu-

nities. The previous study suggested two groups of users in Q-A communities, popular and casual users [139]. This observation motivated us to closely analyze the network segmentation in the core-periphery structure. Based on Bayesian Stochastic modeling, we identify each 30-day network core user. Furthermore, using the DIBRM model[67], 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 active one, but as time progressed, this changed in favor 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 behavior that stays universal. The specific interactions among elements could lead to different kinds of organizational patterns. This thesis aims to understand the factors that drive the system's growth and change its structural properties and 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 describes 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; however, empirical analysis of various real systems and agent-based simulation [39, 40] have indicated that properties of growth signal influence the dynamics of complex systems, as well as the structure of its interaction network. To investigate the connection between the features of the growth signal and the structure of an evolving network, we added one more parameter in the growth of the aging network model, the fluctuating growth signal, and examined how network properties change with the signal features. 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 temporal correlations. The structure of the generated complex networks depends on the signal properties, and 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 [140]. The number of citations scales as lognormal distribution [21]. 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 3: 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. One option is that the user joins a group where she already has friends; it's determined with affiliating probability p_{aff}, while with probability $1 - p_{aff}$, the user chooses a random group. Group size distribution in this model is lognormal. The width of the lognormal distribution depends on the probability p_{aff}; it becomes broader with a larger probability p_{aff}.

In chapter 5, we focused on the factors that influence the sustainability of evolving complex networks. Specifically, we investigated the sustainability of social groups 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. 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 modeling approach [66] 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 [67] (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 question 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 modeling 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.

The results from this thesis contribute to our knowledge about the structure and dynamics of evolving complex networks and how they are mutually linked. We explored different factors that influence network growth, structural properties, and sustainability. The growth signal impacts the network's structural properties, while social interactions affect group segmentation. The sustainability of evolving networks depends on core-periphery structure, the core's stability, and users' ability to form a trustworthy core. Research presented in this thesis confirms that dynamics is linked with the structure of its interaction network, while the structure directly determines the function, organization, and sustainability of complex systems.

Complex network theory is a rapidly growing field, but many open research questions exist. With the increase in the availability of the data of various complex systems, the analysis of complex networks

becomes even more popular and shows excellent potential for future work. While we mostly understand how to describe the network's structure, and many methods are adapted to deal with evolving complex networks, we still need insights into how to design networks in order to control their properties, prevent epidemic outbreaks, and enhance or diminish information diffusion. Incorporating spatial or temporal constraints in network models could provide a more accurate picture of systems evolution. Community detection methods are beneficial for understanding network structure and function, but it lacks methods that easily adapt to network changes over time. The current development of deep learning on graphs could fill existing gaps and provide more accurate predictions of complex network systems' behavior.

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 [141], 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 and 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%

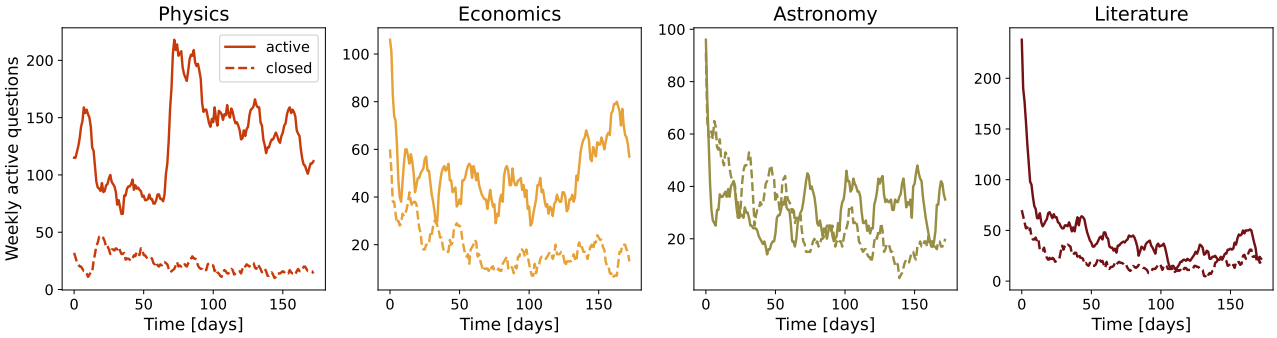

Figure A.1: Number of active questions within seven days sliding windows. Solid lines - active sites; dashed lines - closed sites.

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 [67, 142] was used to approximate real reputation on Stack Exchange sites [142], 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.

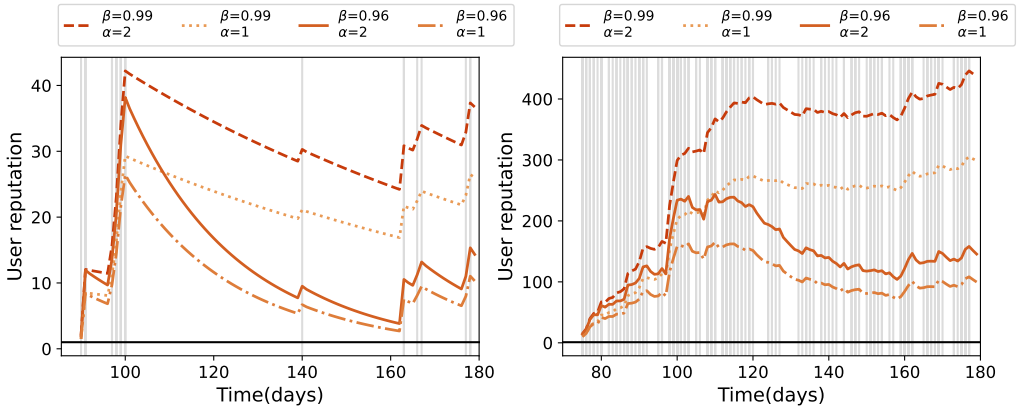

Figure B.1: Single users reputations, left panel shows sporadically active user, while user on right makes frequent interactions.

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.

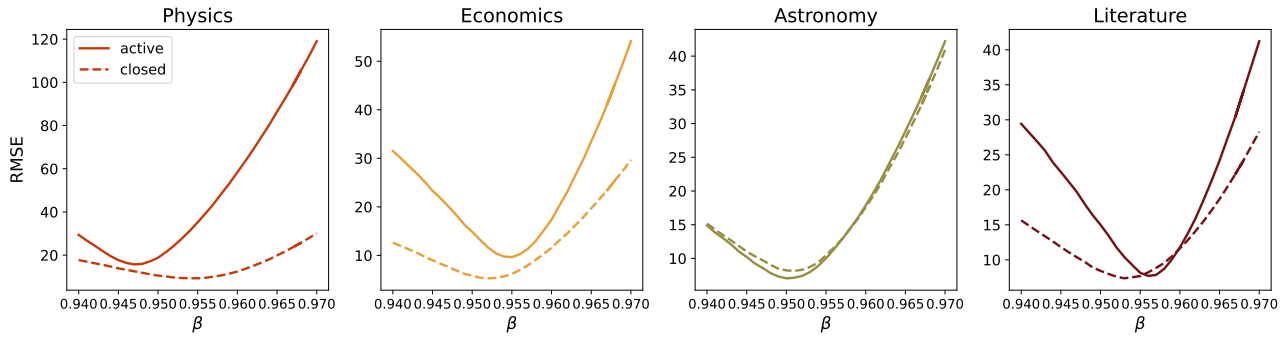

Figure B.2: RMSE between the number of active users in a sliding window of 30 days and the number of users with reputation > 1 for $0.94 < \beta < 0.97$ with step 0.001.

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.

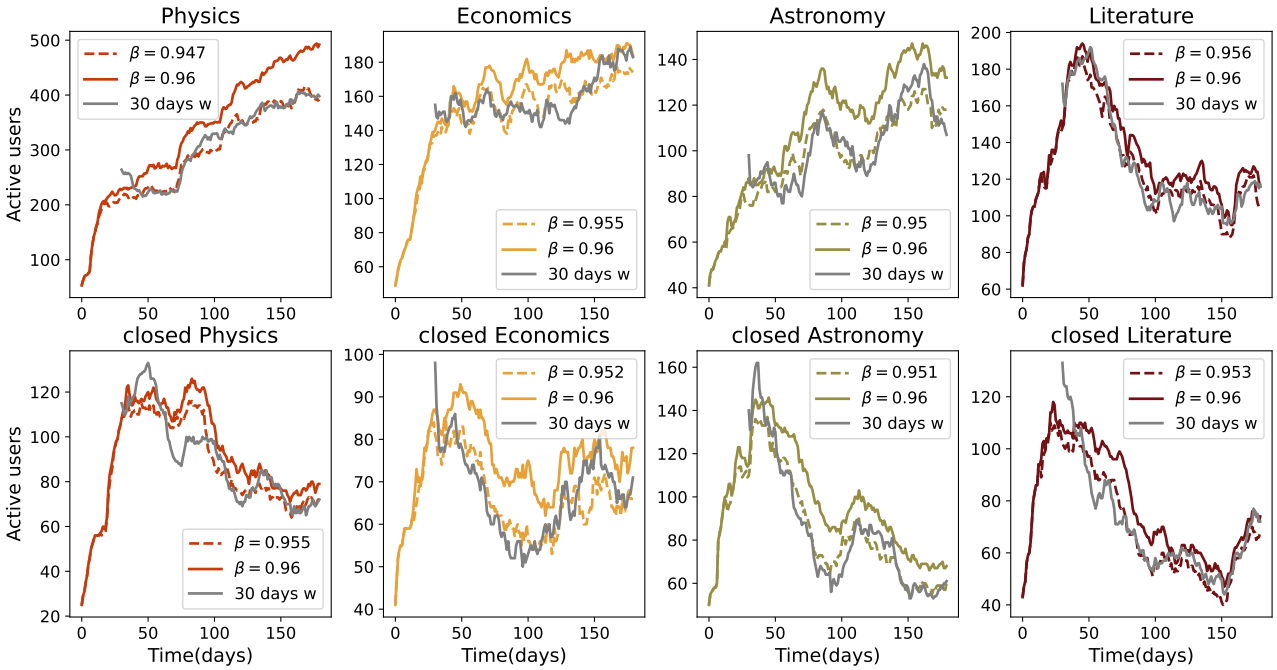

Figure B.3: Number of active users in a sliding window of 30 days and number of users with dynamic reputation higher than 1 for $\beta = 0.954$ and $\beta = 0.96$ which provide the best fit to the number of users in 30 days sub-networks for each community.

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.

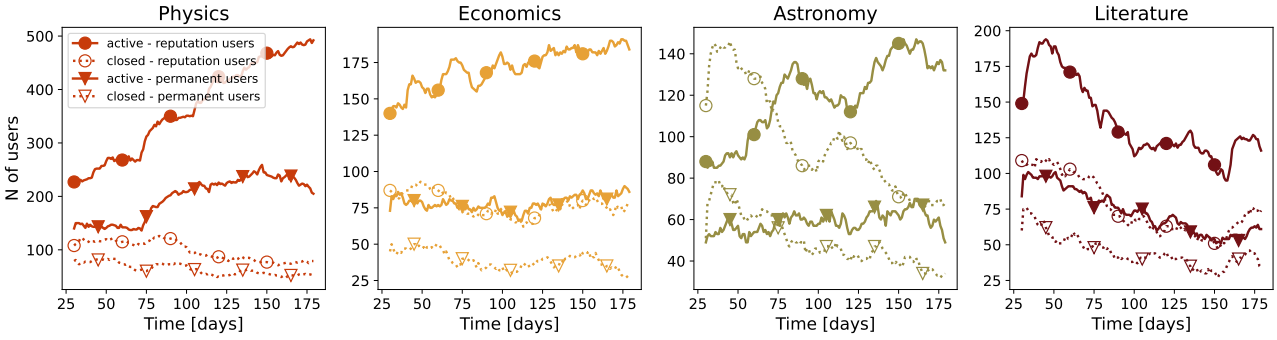

Figure B.4: Solid lines represent the number of users with dynamic reputation higher than 1 for $\beta = 0.96$ while dashed lines are the number of users within 30 days sliding window who were active and remained to be active.

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 [57, 58]. 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.

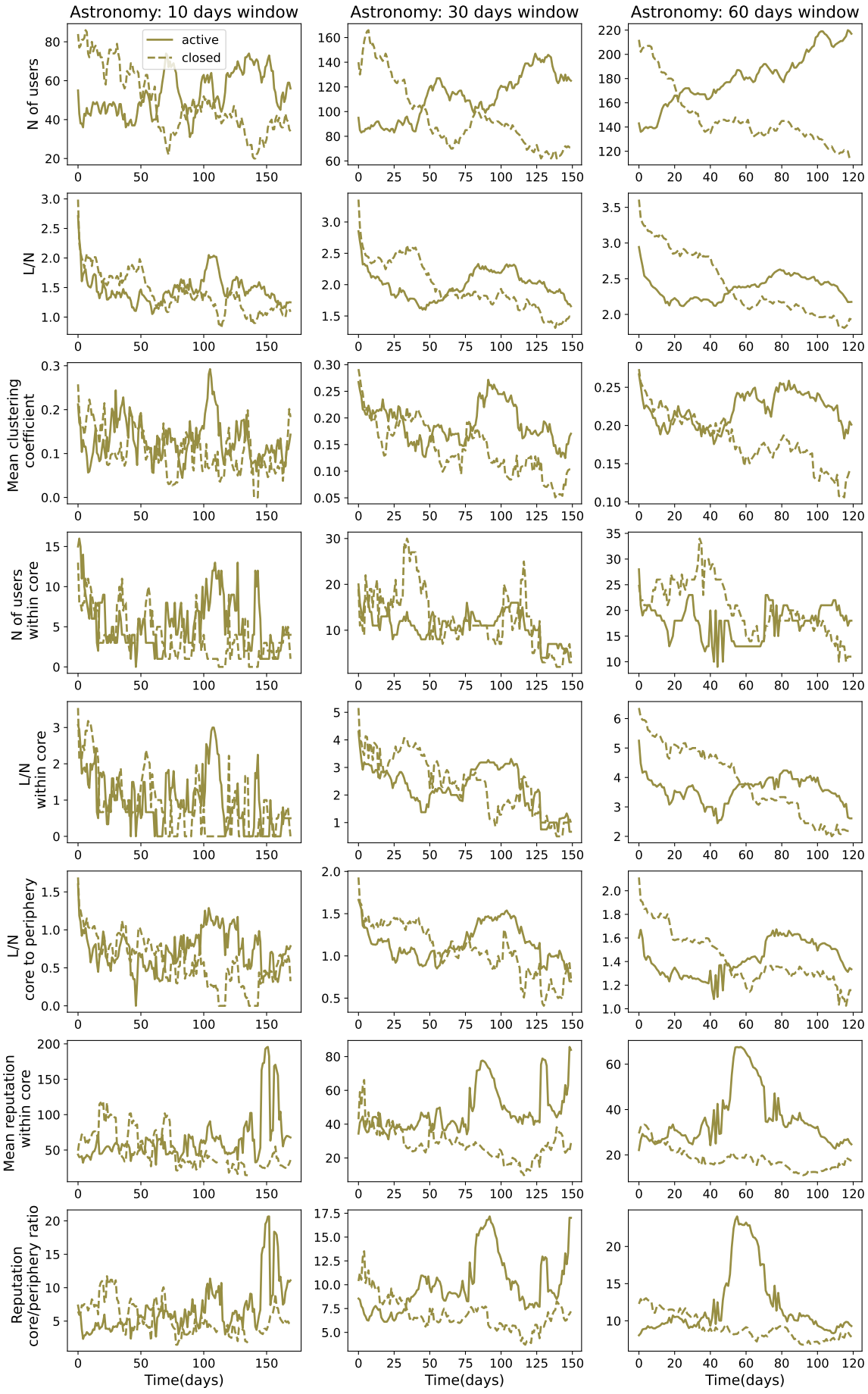

Figure C.1: Results for different sliding windows. For astronomy, solid blue lines- active, orange dashed lines - closed site.

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 [143] 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) [143].

The **F1 measure** is the harmonic mean of precision and recall [143]:

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

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 [143]. 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 [144]:

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

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 [145] 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 [146], 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.

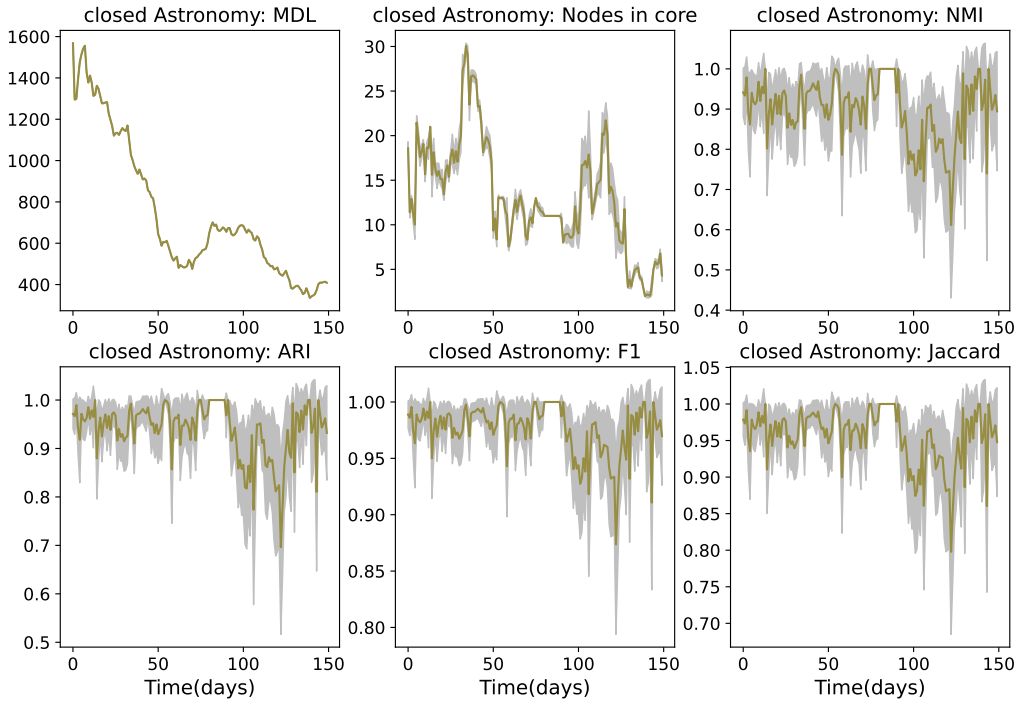

Figure D.1: Minimum description length, number of nodes in the core, normalized mutual information, adjusted rand index, F1 measure and Jaccard index, among 50 samples for 30-days sub-networks. Results are given for closed astronomy.

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] 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.

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

[5] 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.

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

[7] 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.

[8] 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.

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

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

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

[12] 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.

[13] Parongama Sen and Bikas K Chakrabarti. *Sociophysics: an introduction*. Oxford University Press, 2014.

[14] Frank Schweitzer. Sociophysics. *Phys. Today*, 71(2):40, 2018.

[15] 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.

[16] James P Sethna. *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press, USA, 2021.

[17] Leo P Kadanoff. Scaling and universality in statistical physics. *Physica A: Statistical Mechanics and its Applications*, 163(1):1–14, 1990.

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

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

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

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

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

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

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

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

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

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

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

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

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

[35] 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.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

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

[43] 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.

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

[45] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.

[46] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[47] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[48] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*, pages 1775–1784, 2018.

[49] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. 2019. *arXiv preprint arXiv:1902.06673*, 1902.

[50] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.

[51] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

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

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

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

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

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

[57] 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.

[58] 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.

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

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

[61] 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.

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

[63] 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.

[64] Ana Vranić, Jelena Smiljanić, and Marija Mitrović Dankulov. Universal growth of social groups: empirical analysis and modeling. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(12):123402, 2022.

[65] Ana Vranić, Aleksandar Tomašević, Aleksandra Alorić, and Marija Mitrović Dankulov. Sustainability of stack exchange q&a communities: the role of trust. *EPJ Data Science*, 12(1):4, 2023.

[66] 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.

[67] 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.

[68] Ernesto Estrada and Philip A Knight. *A first course in network theory*. Oxford University Press, USA, 2015.

[69] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.

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

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

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

[73] Angélica Sousa da Mata. Complex networks: a mini-review. *Brazilian Journal of Physics*, 50:658–672, 2020.

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

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

[76] 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.

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

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

[79] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.

[80] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science advances*, 3(5):e1602548, 2017.

[81] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.

[82] 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.

[83] 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.

[84] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and blockmodeling*, pages 289–332, 2019.

[85] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

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

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

[88] Mark EJ Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004.

[89] 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.

[90] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.

[91] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.

[92] 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.

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

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

[95] 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.

[96] 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.

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

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

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

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

[101] 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.

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

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

[104] 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.

[105] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.

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

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

[108] 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.

[109] 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.

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

[111] 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.

[112] 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.

[113] 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.

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

[115] Albert-Laszlo Barabâsi, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.

[116] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.

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

[118] 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.

[119] 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.

[120] 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.

[121] 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.

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

[123] 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.

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] 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.

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

[130] 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.

[131] 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.

[132] János Török and János Kertész. Cascading collapse of online social networks. *Scientific reports*, 7(1):1–8, 2017.

[133] Thilo Gross and Bernd Blasius. Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008.

[134] Xiao Han, Shinan Cao, Zhesi Shen, Boyu Zhang, Wen-Xu Wang, Ross Cressman, and H Eugene Stanley. Emergence of communities and diversity in social networks. *Proceedings of the National Academy of Sciences*, 114(11):2887–2891, 2017.

[135] László Lőrincz, Júlia Koltai, Anna Fruzsina Győr, and Károly Takács. Collapse of an online social network: Burning social capital to create it? *Social Networks*, 57:43–53, 2019.

[136] 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.

[137] 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.

[138] 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.

[139] 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.

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

[141] 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.

[142] 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.

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

[144] 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.

[145] 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.

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

Biography of the author

Ana Vranić was born on November 23rd, 1993, in Čacak, Republic of Serbia, where she finished elementary and high school. In 2012 she enrolled BSc studies of Theoretical and Experimental Physics at the Faculty of Physics Belgrade and graduated in 2016 with a GPA of 9.24/10.00. In the same year, she started MSc studies at the Faculty of Physics and, after one year, finished them with a GPA of 10.00/10.00. Her master thesis, "Thermodynamics and electronic transport in Habard model on the triangular lattice", was done under Dr. Darko Tanasković in Scientific Computing Laboratory at the Institute of Physics Belgrade. During this research, she visited the institute Jožef Stefan in Ljubljana, for which she received the CEEPUS scholarship. Ana also won the "Prof. dr Ljubomir Ćirković" foundation award for best MSc thesis defended at the Faculty of Physics of the University of Belgrade.

In 2017, Ana Vranić started PhD studies at the Faculty of Physics in statistical physics. Under the supervision of Dr. Marija Mitrović Dankulov at the Institute of Physics Belgrade. Since April 2018 Ana has been employed at the Institute of Physics Belgrade as a Research Assistant in the Scientific Computing Laboratory of the National Center of Excellence for the Study of Complex Systems. She participated in several projects: the National Project ON171017 Modeling and Numerical Simulations of Complex Many-Body Systems, funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia; Artificial Intelligence Theoretical Foundations for Advanced Spatio-Temporal Modelling of Data and Processes (ATLAS) project funded by the Science Fund of the Republic of Serbia and in REmote development of Autonomous Driving algorithms in a realistic environment (READ) project funded by Innovation Fund of Republic Serbia.

Ana Vranić has published four papers in peer-reviewed international journals. Papers (1-3) are part of this thesis, while the 4th paper presents research done during MSc studies. She also presented her research at several international conferences.

1. **Vranić A**, Tomašević A, Alorić A, Mitrović Dankulov M. Sustainability of Stack Exchange Q&A communities: the role of trust. EPJ Data Science. 2023 Feb 24;12(1):4.

2. **Vranić A**, Smiljanić J, Mitrović Dankulov M. Universal growth of social groups: empirical analysis and modeling. Journal of Statistical Mechanics: Theory and Experiment. 2022 Dec 7;2022(12):123402.

3. **Vranić A**, Mitrović Dankulov M. Growth signals determine the topology of evolving networks. Journal of Statistical Mechanics: Theory and Experiment. 2021 Jan 22;2021(1):013405.

4. **Vranić A**, Vučičević J, Kokalj J, Skolimowski J, Žitko R, Mravlje J, Tanasković D. Charge transport in the Hubbard model at high temperatures: Triangular versus square lattice. Physical Review B. 2020 Sep 21;102(11):115142.

Изјава о ауторству

Име и презиме аутора – **Ана Вранић**

Број индекса – **2017/8006**

Изјављујем

да је докторска дисертација под насловом

Evolving complex networks: structure and dynamics

(Растуће комплексне мреже: структура и динамика)

- 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, 16.03.2023.

Потпис аутора

Ана Вранић

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора – **Ана Вранић**

Број индекса – **2017/8006**

Студијски програм – Физика кондензоване материје и статистичка физика

Наслов рада – **Evolving complex networks: structure and dynamics**

Растуће комплексне мреже: структура и динамика)

Ментор – **др Марија Митровић Данкулов**

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, 16.03.2023.

Потпис аутора

Ана Вранић

Изјава о коришћењу

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

Evolving complex networks: structure and dynamics

(Растуће комплексне мреже: структура и динамика)

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 (Creative Commons) за коју сам се одлучила.

1. Ауторство (CC BY)
2. Ауторство – некомерцијално (CC BY-NC)
3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)
4. **Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)**
5. Ауторство – без прерада (CC BY-ND)
6. Ауторство – делити под истим условима (CC BY-SA)

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

У Београду, 16.03.2023.

Потпис аутора

Ана Вранић

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.