Chapter 1

Methodology

1.1 Complex networks

Many real systems are composed of a large number of elements interacting with each other. Due to interactions, without any central force, the system exhibits the emergence of collective behaviour on the macro level. Such a system is called a Complex System and its properties can not be predicted from the behaviour of the one individual. An example of a complex system is the human brain. The structure of the brain network and its properties are fundamental for brain functioning, while an emergent phenomenon is a human intelligence. In societies, people's interactions lead to civilization, economy, formation of social groups. Also, the animal populations show different levels of organization that emerge from the individual's interactions [1].

The research in complex systems focuses on the structure of the interactions between units. Knowing how branches of the system are connected, we can determine the emergence of the collective behaviour of the system. For the brain network, we can construct representation with neurons and synapses, representing the brain connectivity. Neurons in the same brain area are closely connected [2]. Similarly, we can define communication between people. The structure of these interactions gives us insights, for example, how information propagates through the system. The presence of people with many connections can lead to faster information flow.

Despite the differences between complex systems, they can be studied using complex networks; with sets of nodes (vertices) and links (edges). Elements in the system are nodes, while interactions between them are given as edges. This approximation allows us to treat equally social (graph of actors), biological (network of proteins) or even technological systems (internet, traffic) [3, 4]. In recent years, complex network theory has application in different fields, and the availability of big data incurs its development.

The complex network theory originates from the graph theory in mathematics. The first mathematical problem solved using graph theory was $Konigsberg$ problem of seven bridges. The city $Konigsberg$ had seven bridges connecting the city's parts across the river and the island in the middle. The question was, is it possible to find a walk that crosses all seven bridges only once. Representing the problem as a graph, as in figure 1.1, Euler managed to simplify the problem; the parts of the land are represented as nodes while bridges between them are links. Crossing each bridge only once is possible if each part of the land has an even number of connections. By this it is possible to

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

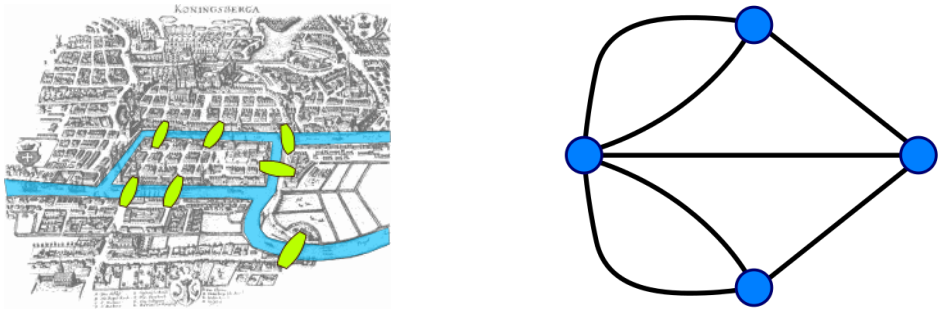

Figure 1.1: The Kronigsber problem of seven bridges.

1.2 Types of networks

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

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

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

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

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

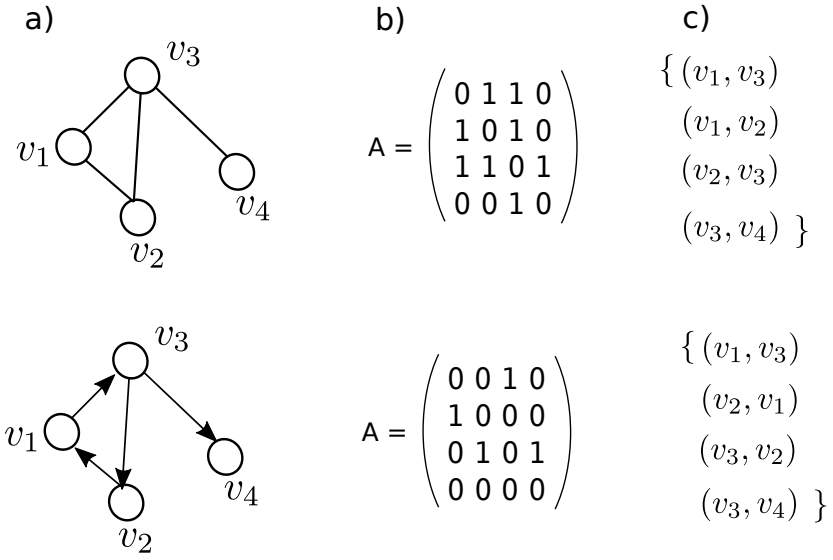

Figure 1.2: a) Graph representation of undirected (top panel) and directed (bottom panel) network. The same networks are represented with adjacency matrices column b), and edge list representation in column c).

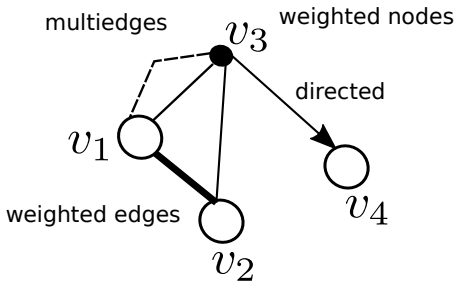

Figure 1.3: The complex networks may represent different characteristics of the system. The edges can be directed, weighted or multiply. Also nodes can be assigned with different weights or any relevant feature.

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

A **bipartite network** consists of two types nodes. The nodes in the same partition are not connected, while links exist only between partitions. For many real systems, a bipartite graph is a natural representation[6, 2]. For example, the bipartite network of people and groups has two distinct node

partitions while links indicate the memberships. Another example is a system of customers and products. The link between user and item is created when the user buys an item. The bipartite networks find their application in the algorithms for recommended systems, whose goal is to recommend items that may interest the user. Actually, to find the most probable missing links in the network.

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

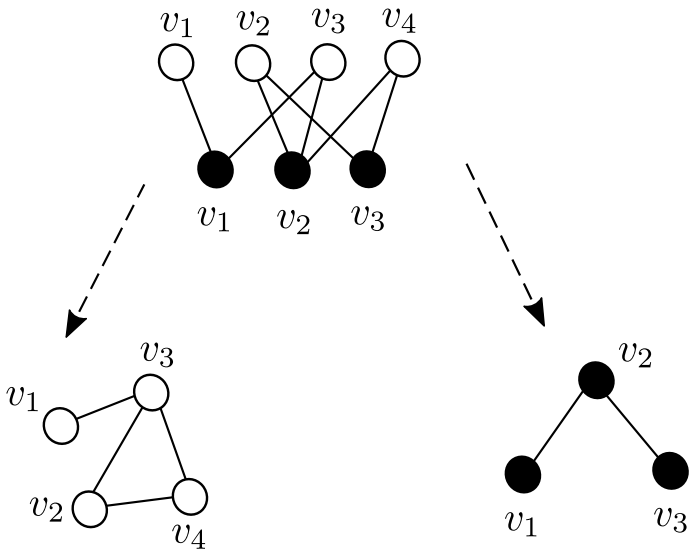

Figure 1.4: Bipartite graph and two partition projections.

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

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

The temporal network is a collection of timestamped edges. Each edge is defined as $(v_i, v_j, t, \Delta t)$, where v_i and v_j, are nodes t is time when interaction happen, and Δt is event duration [9]. The duration of the events may vary, as in the phone-call network. Also, for many systems, the time

resolution of event duration is too small. For example, this parameter may be neglected when people interact on social platforms or email each other because the event time is too short, it scales in seconds.

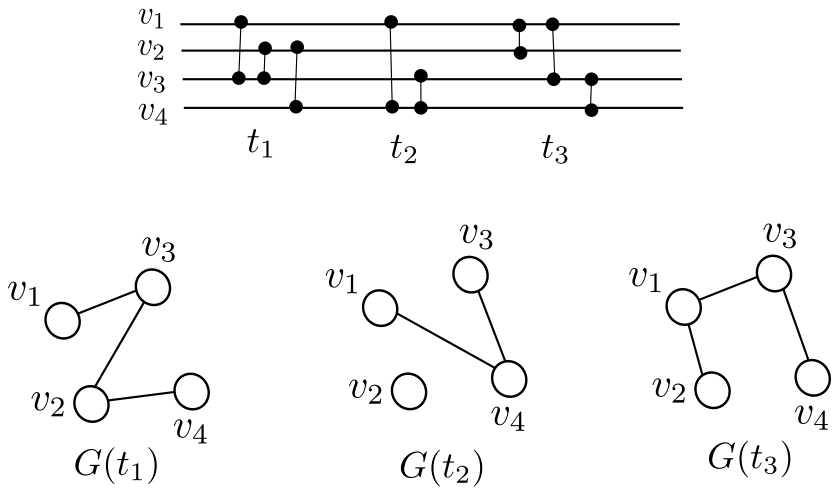

Figure 1.5: Temporal network.

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

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

1.3 The structure of complex networks

1.3.1 Degree distribution

The simplest network measure is **node degree**, k. The degree of node i gives the number of nodes attached to node i, $k_i = \sum_j A_{ij}$. The density of the network is average degree divided by $N - 1$, where N is number of nodes. It is relative fraction of nodes in the network.

In the case of regular networks, such as grids, each node has an equal degree, meaning that nodes in the network have similar roles. In the general case, the networks have more complicated structure.

If degree sequence is skewed, we are able to identify nodes with high degree, hubs. Removing hubs may partition a connected network into several components. Finally, if we are able to test isomorphism between two graphs, the starting point would be to compare their degree sequences are the same. If they are not same, then graphs can not be isomorphic.

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

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

- The Poisson distribution. The degree distribution in random network, where all nodes have the same connecting probability, follows Poisson distribution $P(k) = \frac{(Np)^k e^{-Np}}{k!}$, where k is the mean degree distribution.
- Exponential distribution. $P(k) = e^{-k/k}$. This is degree distribution of the growing random graph. Even for infinite networks all moments of distributions are finite, and have natural scale of the order of average degree.
- In many real networks degree distribution follows a power law. $P(k) = k^{-\gamma}$, where γ is exponent of the distribution. In this distribution there is no natural scale, so they are called scale-free networks. In infinite networks all higher moments diverge. If the average degree of scale-free networks is finite, than γ exponent should be $\gamma > 2$. Therefore, real networks have a scale-free structure with the emergence of the hubs [7].

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

1.3.2 Degree correlations

Correlation is defined through a correlation coefficient r. If x and y are two stochastic variables, for which we have a series of observation pairs $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$. The correlation coefficient $r(x, y)$ between x and y is defined as:

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

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

Taking the definition of correlation coefficient we can define it for vertex degrees. For simple graph G with vertex set $V(G) = \{v_1, ..v_n\}$, $\boldsymbol{A}[i,j] = 1$ if there is a link between nodes v_i and v_j. If G is a simple graph with adjacency matrix $\boldsymbol{A}$ and degree sequence $\boldsymbol{d} = [d_1, ..., d_n]$

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

Using adjacency matrix, allow us to calculate the correlations between neighboring nodes. If two nodes are not connected $A[i,j] = 0$, the degree correlation between them does not have contribution to the r.

The **degree-degree correlations** in the network are measured by **assortativity**. If correlations are positive, networks are assortative; there is a tendency that connections exist between similar degree nodes. The negative correlations indicate that large degree nodes have preference to connect nodes with small degree; dissasortative networks. The average first neighbor degree k_{nn} can be calculated as $k_{nn} = \sum_{k'} k^{'} P(k^{'}|k)$. The P is conditional probability that an edge of degree k points to node with degree k. The norm is $\sum_{k'} P(k^{'}|k) = 1$, and detailed balance conditions [1], $kP(k^{'}|k)P(k) = k^{'}P(k|k^{'})P(k^{'})$ [1]. If the node degrees are uncorrelated, k_{nn} does not depend on the degree, otherwise increasing/decreasing function indicates on positive/negative correlations in the network.

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

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

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

1.3.3 Clustering coefficient

The **clustering coefficient** is a measure describing the neighbourhood's structure. In networks exist tendency to form triangles or clusters. This is common in friendship networks where two friends of one person have a high probability of being friends. The clustering can be measured by computing the number of links between neighbours of one node,

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

Averaging it over all network nodes, we can calculate the mean clustering coefficient. It ranges from $\langle c \rangle = 0$ where connections between neighbouring nodes do not exist, network has the structure of three. On the other hand, $\langle c \rangle = 1$ indicates a fully connected network.

Newman proposed the alternative definition for the clustering coefficient based on the number of triples and triangles in a graph. A triangle at node v is complete subgraph with 3 nodes, including v. A triple on the node v is a subgraph of exactly three nodes and two edges, where v is incident with two edges. The network transitivity is defined as the ratio of number of triangles in the network over the number of triples. The network transitivity is seen as global clustering, as it considers the whole network.

1.3.4 Network paths

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

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

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

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

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

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

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

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

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

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

1.3.5 D-measure

For each node i we can define the distribution of the shortest paths between node i and all others nodes in the network, $P_i = \{p_i(j)\}$, where $p_i(j)$ is percent of nodes at distance j from node i. The connectivity patterns can efficiently describe difference between two networks. To specify how much G and $G^{'}$ are similar we use D-measure [16]

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

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

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

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

The first term in equation 1.8 compares local differences between two networks, and Jensen-Shannon divergence between N shortest path distributions $J(P_1, ..., P_N)$ is normed with network diameter $d(G)$. The second part determines global differences, computing $J(\mu_G, \mu_{G'})$ between mean shortest path distributions. The D-measure ranges from 0 to 1. The lower D-measure is, networks are more similar and for D-measure $D = 0$, structures are isomorphic.

1.3.6 Community structure

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The another way to detect core-periphery structure is to use the inference method based on fits to a stochastic block model. In this method we fit observed network to a block model with two groups, such that edge-probabilities have form $p_{11} > p_{12} > p_{22}$.

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

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

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

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

Prior on p is modified for core-periphery model such that $P(p) = 3! I_{0<p_{22}<p_{12}<p_{11}<1}$, while prior on θ consists of three parts: probability of having l blocks; given the number of layers probability

$P(n|l)$ of having groups of sizes $n_1..n_l$ and the probability $P(\theta|n)$ of having particular assignments of nodes to blocks.

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

1.4 Network models

1.4.1 Random network model

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

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

- we start with N isolated nodes
- between each $N(N-1)/2$ pair of nodes we create link with probability p; sampling random number $r \in (0, 1)$, we create link if $r \leq p$

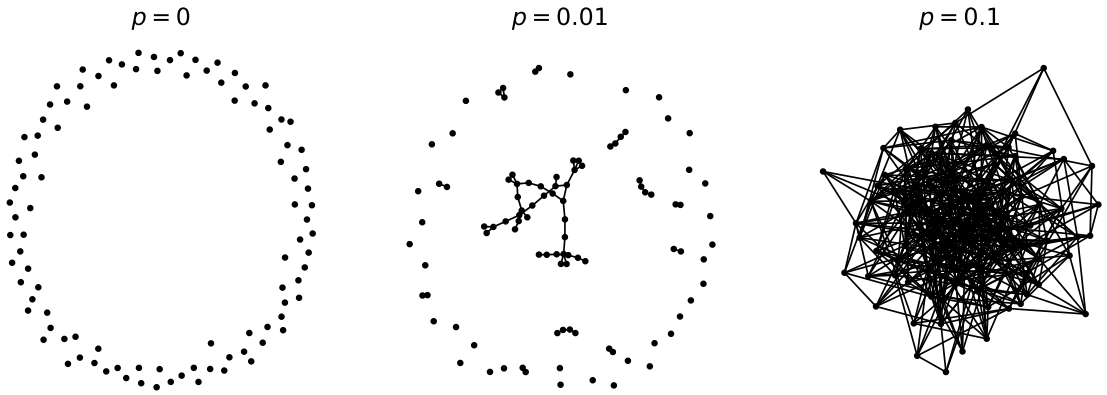

Figure 1.6: ER graph with $N = 100$ nodes and different linking probabilities p.

We should note that this process is stochastic. The networks $G(N, p)$ with the same parameters do not need to have the same structure; i.e. they differ in the number of links. Therefore, the single random graph is only one graph from all the possible realizations in the statistical ensemble.

Two simple quantities that could be estimated are the average number of links and the average degree. For complete graph with N nodes, number of edges is $N(N-1)/2$. As the probability of drawing every edge is p, the **average number of links** is simply given as

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

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

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

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

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

The probability that the node has degree k is given with the second term p^k, while the probability that other N-1-k links are not created is given with the third part of the equation. Finally, there are $\binom{N-1}{k}$ combinations for one node, to have k links from $N-1$ possible links.

The binomial distribution describes very well small networks. For larger networks, we find that they are sparse and that the average degree is much smaller than a number of nodes $\langle k \rangle << N$. In this limit, binomial distribution becomes the Poisson, which now depends only on one parameter $\langle k \rangle$

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

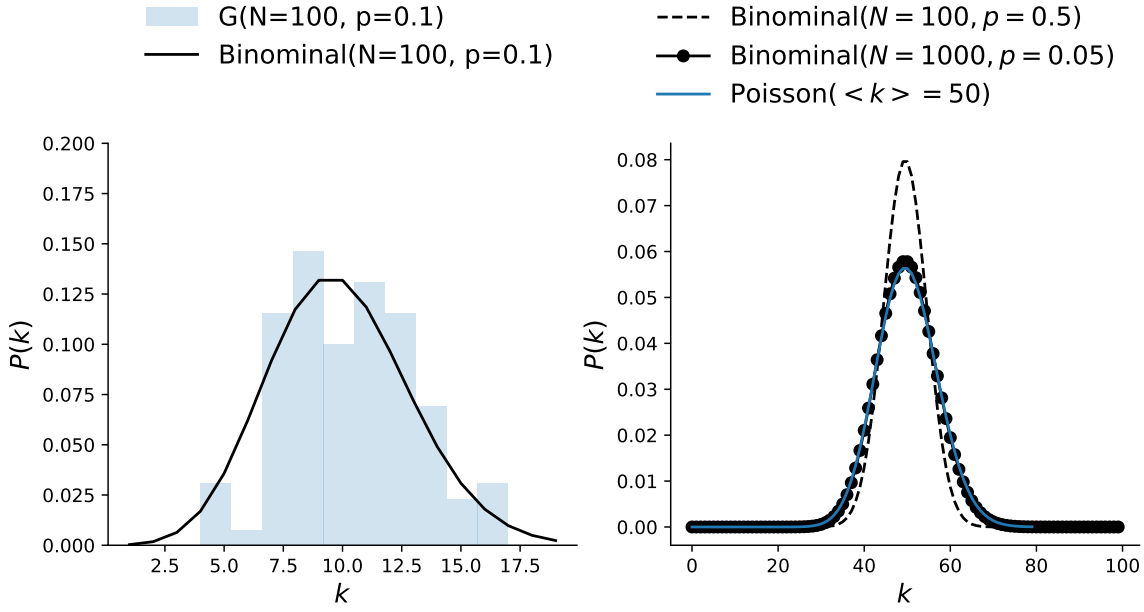

Figure 1.7: Degree distribution of ER graph. Degree distribution of small networks follow binominal. Larger networks are better approximated with Poison distribution, and degree distribution for fixed average degree $< k >$ becomes independent of the network size.

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

The figure 1.6 shows how the network becomes more connected by increasing the linking probability p. When $p = 0$, all nodes are disconnected. In the other limit, $p = 1$, the network is fully connected. Between those two probabilities exists critical probability, where the giant component appears. The giant component is a sub-graph, which size is proportional to the network size. In other words, the network does not have disconnected components. Such change in the network is a phase transition in network connectivity and is related to percolation theory.

The phase transition occurs when average degree is $\langle k \rangle = 1$, which gives us: $p_c = \frac{1}{N-1}$, meaning that all nodes have degree larger than 1. When the $\langle k \rangle < 1$, the network is in the sub-critical regime where all components are small. In the critical regime, the size of the giant component is proportional to the $N^{2/3}$. In the supercritical regime, $\langle k \rangle > 1$, the probability of a giant component appearing is 1.

1.4.2 Small-world networks

Inspired by the idea that real-world networks are highly clustered and the average distance is small, Watts and Strogatz proposed the "small-world" model. The model starts from the regular lattice, and with rewiring links, the network starts to resemble small-world property. The procedure is the following:

- At the beginning, nodes are placed on the ring lattice, and each node is connected to $k/2$ first neighbours on the left and the right side. Initially, the clustering coefficient is high, $c = 3/4$.
- For each link in the network, with probability p, we choose a random node to rewire the link. This makes long-distance nodes connect, decreasing the network's average path length.

The model interpolates between the regular graph when the probability is $p = 0$ and the random graph with $p = 1$ when all links are randomly rewired. Short distances and high clustering are present in the network for the critical probabilities.

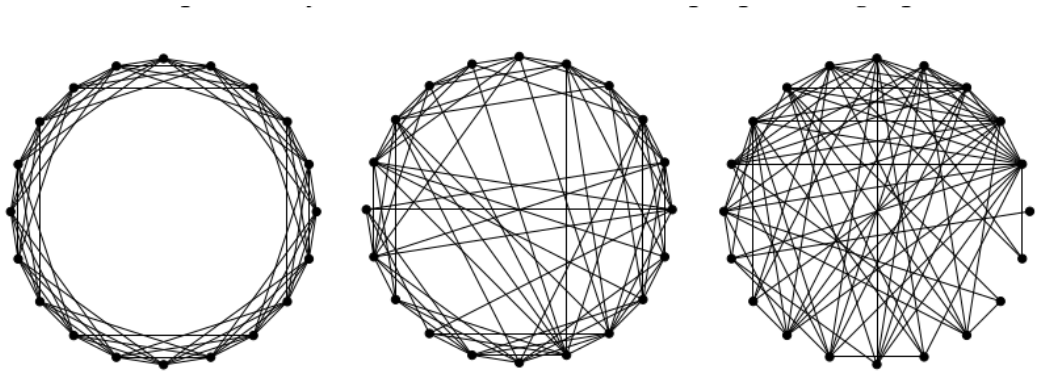

Figure 1.8: Watts and Strogatz graph model creation

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

1.4.3 Barabási-Albert model

The ER random graph model and WS small-world model are static models, where the number of nodes is fixed. It is one of the reasons why they can not fully explain the properties of real systems. The size of real systems does not remain constant; real networks grow. For the network, the growth means that at each time step, new nodes are added to the network. The simplest model that produces the scale-free networks is Barabasi-Albert model.

- The model starts from the small number, n_0 randomly connected nodes, with m_0 links.

- At each time step, new node with m links joins to the network. New node creates links with the nodes already present in the network, following the linking rules; i this case rules of preferential attachment.

The preferential attachment is important ingredient for generating system with scale-free properties. In the real-system the linking between nodes is not random process, there exists the preference toward specific types of nodes. For example the popular web-pages can easily get more visits or it is common that already popular papers will get more citations. This effect is also called rich-get-richer or preferential attachment.

The simplest formulation of the preferential attachment model is that new nodes tend to connect with high degree nodes. The linking probability Π is then proportional to node degree k:

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

As at each time step one node arrive, we can estimate the number of nodes at the time step t, $N(t) = n_0 + t$, with links $L(t) = m_0 + mt$.

First we can calculate the evolution of network degree in time.

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

Note that new node, that arrived at time point t_i has degree m, as it links to m old nodes. Solving the equation we get that at $t > t_i$, has degree that grows as square root of time, also it shows that younger nodes easily acquire larger degree.

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

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

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

For large k it is exactly power-law. It is also independent of the time and size of the system, meaning the emergence of stationary scale-free state. Distributions do not depend on the N. If we vary m the slope of distributions is the same, but they are parallel. After re-scaling $p(k)/m^2$, they fall on the same line.

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

The combination of the growth and preferential attachment linking is crucial for getting scale free networks. For example, eliminating the preferential attachment; in growing network with random

linking, degree distribution is stationary, but it follows exponential. In contrast, the absence of growth leads to the non-stationary degree distribution. When number of nodes is fixed, while the network grows only in number of links, such that randomly chosen node i connects to node j according to probability Π. At the beginning, the degree distribution follows the power-law, same as in BA model. As more links are added to the network, the distribution changes it's shape, first the peak appears, while at the end network becomes complete graph, where all nodes have the same degree.

1.4.4 Nonlinear preferential attachment model

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

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

The probability that newly added node attaches to node i depends on the existing i-th node degree k_i, and the parameter β. When $\beta = 1$, the model is BA model, where degree distribution follows the power-law. When $\beta = 0$, linking probability becomes uniform; i.e. it corresponds to random network model, and degree distribution is Poisson; there is exponential decay.

For $\beta > 1$, the effects of preferential attachment are increased, leading to emergence of super-hubs. The hub-and-spoke network appear in this regime, where almost all nodes are connected to few high-degree nodes.

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

1.4.5 Ageing model

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

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

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

If $\alpha > 0$ is positive, the older nodes are more likely to create connections. In this regime, the preferential attachment stays present, and the high-degree and older nodes are preferred. For very high α, each node is connected to the oldest node in the network. The scale-free properties are present; the power-law exponent γ deviates from $\gamma = 3$. It is found that γ ranges between 2 and 3.

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

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

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

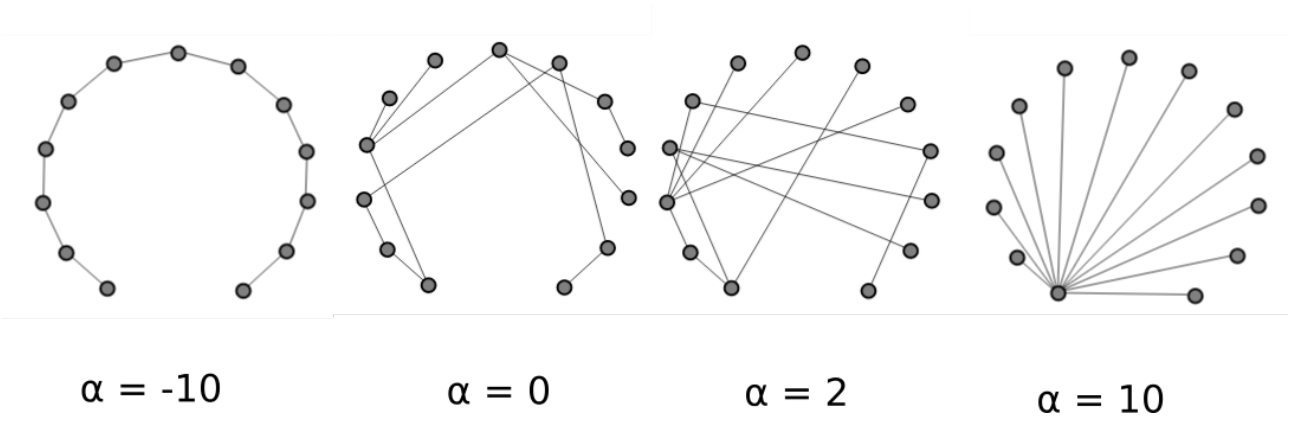

Figure 1.9: Aging model

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

- For example if we fix $\beta = 1$ and $\alpha = 0$ generated networks are scale-free; degree distribution is $P(k) \sim k^{-\gamma}$ with $\gamma = 3$.
- In the case of nonlinear preferential attachment $\beta \neq 1$ and $\alpha = 0$ scale-free properties disappear.
- Scale-free property can be produced along the critical line $\beta(\alpha^*)$ in the $\alpha - \beta$ phase diagram, see Figure 1.10.
- For $\alpha > \alpha^*$ networks have **gel-like small world** behavior.
- For $\alpha < \alpha^*$ and near critical line $\beta(\alpha^*)$ degree distribution has **stretched exponential** shape

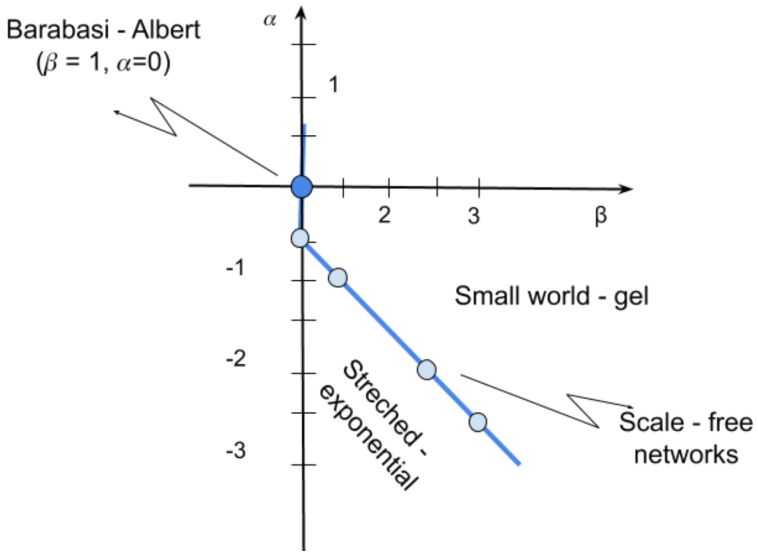

Figure 1.10: Phase diagram of aging network model

1.4.6 Stochastic block model

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

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

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

SBM can generate and describe different types of network structures. Figure 1.11 [17] shows how the model matrix corresponds to resulting networks with two communities. First, for the assortative network (1.11 a), diagonal elements of the matrix have higher probabilities. This indicates dense connections inside the group, just like in classic community structures. In disassortative structure, (1.11 b), more connections exist between two partitions than inside them, i.e. off-diagonal elements have higher probabilities. Bipartite networks can be represented like this.

Figure (1.11 c) shows how the model represents core-periphery networks. Nodes of one block (core) are well connected with itself and with other partition (periphery). From the last case, we can note that SBM with one group is the Erdos Renyi random graph (1.11 d) because all probabilities inside and between groups are equal.

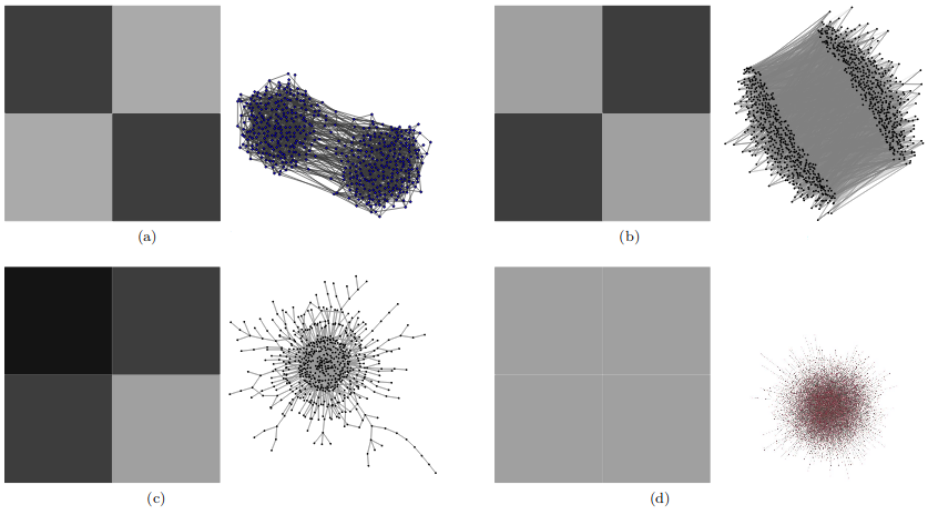

Figure 1.11: Stochastic Block model for different networks structures. (a) assortative. (b) dissortative. (c) core-periphery. (d) Erdos Renyi random graph.

The benefit of this model is that we can generate many networks with similar group structure. The model can fit real data, which results in finding network communities. For the given network G and number of groups k, the best nodes partition g is found by maximizing the likelihood function. Beside inferring communities, SBM has application in prediction of missing links. This simply formulated model has many variants, motivated by specific properties of real data. For example, for networks which are degree heterogeneous, there is degree corrected SBM. In some social networks, users can

belong to more than one group, and this can be modelled with mixed membership SBM. Other extensions include application to bipartite, weighted network, hierarchical model, etc. Also, several algorithms for optimization of likelihood function are proposed. The overview of these versions and methods are given in [21].

1.5 The probability distributions

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

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

This section gives an overview of relevant distributions and methods for fitting data and testing the quality of the performed fit.

1.5.1 The properties of distributions

Power-law The power-law distribution is defined as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Power law with exponential cut-off. The density function has following form

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

where $k > 0$ and $\gamma > 0$. This function combines the power-law and exponential terms responsible for an exponentially bounded tail. Taking the logarithm $ln(p(k)) = lnC - \gamma lnk - \lambda k$, when $k << 1/\lambda$

the second term dominates, so distribution follows te power-law, with exponent γ. Otherwise, the λx term dominates, resulting in an exponential cutoff for high values.

Streched exponential The stretched exponential distribution is defined as:

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

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

1.5.2 Plotting the distributions

The first step in analysis the empirical data is to create the frequency plot, or histogram. Data are binned in the equal intervals and the number of data points within interval are plotted. When plotting heavy tailed distributions it is hard to determine if distribution is for example exponential or power law. To make the distribution visualization more appropriate it is possible to scale the data. If data are from power law distribution on the double logarithmic scale they will look linear:

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

On the log-log scale we can notice that in the tail of the distribution data are noise. As the size of the bins is constant the density of the bins for large values becomes also large. To avoid the fluctuations in the tail, we can use logarithmic binning. The noise is reduced by dividing the x axis into n bins $b_n = c^n$, so the following bin is wider than previous one. For the base c we can choose any value $c > 1$. Similarly, the binning can take the following form $b_n = k_0 \exp(cn)$, where k_0 is the minimum data point, while the c is the arbitrary base. All data points between values $[b_n, b_{n+1})$ are represented with one point $p(k_n) = N_n/b_n$, where N_n is number of nodes found in the bin b_n and $k_n = \sum_i k_i/N_n$ is average degree of the nodes in the bin b_n. By this, averaging over bins in the tail, noise in the tail of distribution is reduced. Still, no matter how bin size is chosen the information about original data points is lost, especially in the tail of distribution where bins are larger and include more samples. Figure 1.12 shows how different distributions look like on linear (first column) and log-log scale (second column).

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

1.5.3 Estimating the distribution parameters

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

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

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

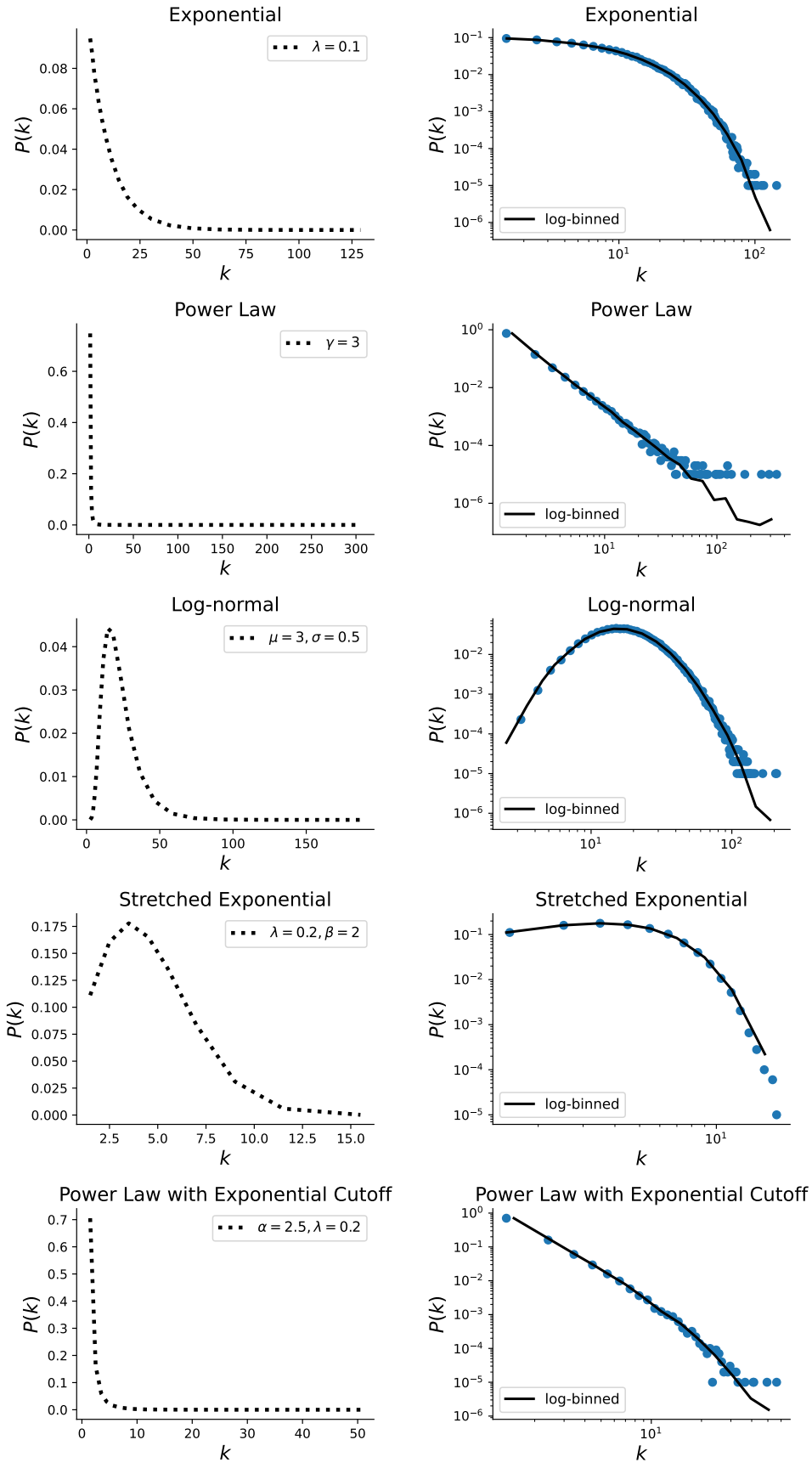

Figure 1.12: Probability distributions on linear and double logarithmic scale.

We can solve the equation and derive the expression for maximum likelihood parameters. For distributions where analytical solution is not available, the parameters can be obtained with numerical optimization. In practice is much easier to work with logarithm of the likelihood, $log(L) = \sum_{i=1}^{i=N} p(\theta|x_i)$, as the product changes to summation. For the power-law distribution the exponent is calculated as $\gamma = 1 + n[\sum ln\frac{k_i}{k_{min}}]^{-1}$. For discrete distribution solution may be obtained optimizing the log-likelihood function $log(L) = log\prod_{i=1}^{n}\frac{k_i^{-\gamma}}{\zeta(\gamma,k_{min})}$.

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

The Kolmogorov Smirnov statistics is the maximum distance between the CDF of the data and the fitted model.

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

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

If $p - value < 0.1$ then 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 correct distribution for the data. There might be other distributions that fit the data equaly good, or even better. For having accurate p-value we need large sample. For small number of syntetic distributions it is possible to have high p-value, even if the distribution is wrong model to the data. Finally, we need to be confident in obtained results. The same procedure can be repeated for different distributions. If p-value for powe law is high, while for alterantive distribution it is low, we can conclude that power-law is more probable fit.

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

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

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

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

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

When $R > 0$ the first distribution is better fit to data and when $R < 0$, the other one should be chosen. When $R = 0$, it is not possible to distinguish between two distributions. The sign of R is not

enough criteria to conclude which distribution is better fit. It is a random variable subject to statistical fluctuations. We need log-likelihood ratio that is sufficiently positive or negative, and to be sure that its sign is not result of fluctuations.

If we are suspected that the true expectation value of the log-likelihood ratio is zero, meaning that observed sign of $\mathcal{R}$ is simply product of fluctuations and an not be trusted. The probability that measured log likelihood ratio has magnitude as large or larger than observed value R is given as

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

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

1.6 Fractal analysis

Approach to study complex systems is detecting time-series of selected variables. Some systems are characterised by periodic ot nearly periodic behaviour. In complex systems this periodic behaviour is not limited to one or two characteristic frequencies. They extend over wide spectrum and fluctuations on many time scales as well as well as broad distributions. In these cases dynamics of the system is characterized by scaling laws, which are valid over a wide range of time scales or frequencies. If dynamic of the system can be described with one scaling exponent system is monofractal, otherwise we deal with multifraactal time-series.

Rescaling of time t by a factor a may require rescaling of the time-series values $x(t)$ by factor a^H, to get the self-similaraity. In this case it is:

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

The Hurst exponent characterize the type of self-affinity.

Many records do not exhibit a simple mono fractal scaling behaviour. The scaling behaviour may be more complicated , and different scaling exponents can be found for many interwoven fractal subsets of the time series. In this case the multifractal analysis must be applied.

Two general types of multifractality exist. The multifractality due to a broad probability distribution for values of the time series, the multifractality can not be destroyed. Multifractality due to different long-term correlations of the small and large fluctuations. In this case the probability density function of the values can be regular distribution with finite moments, and the corresponding shuffled series will exhibit non-multifractal scaling as correlations are destroyed with shuffling procedure. If both kinds of multifractality are present, the shuffled series will show weaker multifractality than the original series. Multifractal analysis will reveal higher order correlations, multifractal scaling can be observed if the scaling behaviour of small and large fluctuations is different. Extreme events might be more or less correlated than typical events.

Long and Short-term correlations The time-series are presistent such that a large value is usually followed by large values and small values. Considering the increments $\delta x_i = x_i - x_{i-1}$, of self-affine series $i = 1, .., N$, with N values measured equidistant in time, so δx_i can be either persistent, independent or anti-persistent. For the random walk with $H = 0.5$ the increments are independent of each other. Persistent and anti-persistent increments, where a positive increment is likely to be followed by another positive or negative increment.

For stationary data with constant mean and standard deviation the auto-covariance function can determine the degree of persistence.

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

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

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

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

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

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

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

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

This type of behavior can be modeled by Fourier filtering techniques. Long-term correlated, behavior of Δx_i leads to self-affine scalling beahviout characterized by $H = 1 - \gamma/2$.

A direct calculation of the $C(s)$ is difficult due to present noise in the data and nonstationarity. Non-stationarities make the definition of $C(s)$ probalematic, because the average is not well defined, also $C(s)$ fluctuates around zero on large scales s, so it is not possible to obtain the correct correlation exponent γ.

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

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

Substracting the averages, constant trends in the data are eliminated. Finally the differences between minimum and maximum value and the standardan deviation in each segment are calculated as:

$$R_\nu(s) = maxY_\nu(j) - minY_\nu(j)$$
$$S_\nu(s) = \sqrt{\frac{1}{s} \sum Y_\nu^2(j)}$$

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

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

the H is Hurst exponent introduced in the first equation. The values of H that can be obtained by Hurst rescaled analasys are $< 0H < 2$. Values $H < 1/2$ indicate long-term anticorrelated data, $H > 1/2$ indicated long-term positively correlated data. For power-lae correlations decaying faster than $1/s$, $H = 1/2$, like for uncorrelated data.

On the other hand the standard fluctuation analysis is based on the random walk theory. For time series with zero mean, we consider the global profile, the cumulative sum:

$$Y(j) = \sum x_i$$

, and then study how fluctuations of the profile, in a given time window of size s increase with s.

We first divide each record of N elements into Ns non-overlaping segments of the size s, and another Ns non-overlaping segments starting from the end. Then we calculate the fluctuations in the each segment. In the standard FA we get the fluctuations just from the values of the profile at both endpoints of each segment.

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

Then we can average F^2 over all subsequences to obtain the mean fluctuation

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

For the relevant case of long-term correlations where $C(s)$ follows the power-law behaviour, $F_2(s)$ also increases by power-law.

The fluctuation exponent is indentical with Hurst exponent for monofractal data.

1.6.1 Detrended fluctuation analysis

DFA was introduced by Peng et al. and represents an important method for describing the non-stationary time series. As before methods it is also base on random walk theory, and the FA analysis is special case with linear detrending.

Like in FA method first one calculates the global profile of the time series. DFA deaks with monotonous trends in a detrending procedure. This is done by establishing a polynomial trend within each segment by least-square fitting and subtracting this trend from the original profile, detrending.

$$Y_s(j) = Y(j) - y^m_{\nu,s}(j)$$

The degree of the polinomial can be varied in order to eliminate constant m=0, linear m=1, quadratic m-2, or higher order trends of the profile function. The variance of the detrended profile in each segment gives us mean-square fluctuations

$$F^2(\nu s) = \frac{1}{s} \sum Y^2_s(j)$$

Finaly fluctioations over all segments are averaged, to obtain the mean fluctuations $F_2(s)$ as in eq. F2s, and as before, from the scaling of the fluctuating function we can determine the Hurst exponent.

1.6.2 Multifractality of the signals

Multifractal detrended fluctuation analysis (MFDFA) [24, 25] to estimate multifractal Hurst exponent H(q). For given time series $\{x_i\}$ with length N, first we define global profile in the form of cumulative sum, equation 2.1, where where $\langle x \rangle$ represents average of the time series:

$$Y(j) = \sum_{i=0}^{j} (x_i - \langle x \rangle), \quad j = 1, ..., N \tag{1.39}$$

Subtracting the mean of the time series is supposed to eliminate global trends. The profile of the signal Y is divided into $N_s = int(N/s)$ non overlapping segments of length s. If N is not divisible with s the last segment will be shorter. This is handled by doing the same division from the opposite side of time series which gives us $2N_s$ segments. 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 detrended signal is calculated as in equation 2.2:

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \left[Y(j) - p^m_{\nu,s}(j) \right]^2 \tag{1.40}$$

Then the q-th order fluctuating function is:

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

The value of $H(0)$, which corresponds to the limit $H(q), q - 0$, cannot be determined directly because of the exponent diverge. Instead logarithmic averaging procedure has to be considered.

$$F_0(s) = \exp \left\{ \frac{1}{4N_s} \sum_{\nu}^{2N_s} ln \left[F^2(\nu, s) \right] \right\}, q = 0 \tag{1.41}$$

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 $H(q)$.

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

Also, large fluctuations are characterized by smaller scaling exponent.

Multifractal signal has different scaling properties over scales while monofractal is independent of the scale, i.e., H(q) is constant.

1.7 Dynamical reputation model

Any dynamical trust or reputation model has to take into account distinct social and psychological attributes of these phenomena in order to estimate the value of any given trust metric [26]. First of all, the dynamics of trust is asymmetric, meaning that trust is easier to lose than to gain. As part of asymmetric dynamics, in order to make trust easier to loose the trust metric has to be sensitive to

new experiences (recent activity or the absence of the activity of the agent), while still maintaining nontrivial influence of old behavior. The impact of new experiences has to be independent of the total number of recorded or accumulated past interactions, making high levels of trust easy to lose. Finally, the trust metric has to detect and penalize both the sudden misbehavior and the possibly long term oscillatory behavior which deviates from community norms.

We estimate dynamic reputation of the Stack Exchange users using Dynamic Interaction Based Reputation Model (DIBRM) [27]. This model is based on the idea of dynamic reputation, which means that the reputation of users within the community changes continuously through time: it should rapidly decrease when there is no registered activity from the specific user in the community (reputation decay), and it should grow when frequent, constant interactions and contributions to the community are detected. The highest growth of user's reputation is found through bursts of activity followed by short period of inactivity.

In our implementation of the model, we do not distinguish between positive and negative interactions in the Stack Exchange communities. Therefore, we treat any interaction in the community (question, answer or comment) as potentially valuable contribution. In fact, evaluation criteria for Stack Exchange websites going through beta testing, described in SI, do not distinguish between positive and negative interactions. The percentage of negative interactions in the communities we investigated was below 5%, see Table 1 in SI. Filtering positive interactions would also require filtering out comments because they are not rated by the community, and that would eliminate a large portion of direct interactions between the users of a community, which is essential for estimating their reputation.

In DIBRM, reputation value for each user of the community is estimated combining three different factors: 1) *reputation growth* - the cumulative factor which represents the importance of users' activities; 2) *reputation decay* - the forgetting factor which represents the continuous decrease of reputation due to inactivity; *the activity period factor* - measuring the length of the period of time in which the change of reputation happened. In case of Stack Exchange communities, the forgetting factor has a literal meaning, as we can assume that past contributions provided by a user are being forgotten by active users as their attention is captured by more recent content.

In line with the the basic dichotomy of reputation dynamics, which revolves around the varying influence of past and recent behavior, DIBRM has two components: *cumulative factor* - estimating the contribution of the most recent activities to the overall reputation of the user; *forgetting factor* - estimating the weight of past behavior. Estimating the value of recent behavior starts with the definition of the parameter storing the basic value of a single interaction I_{b_n}. Cumulative factor I_{c_n} then captures the additive effect of recent successive interactions. The reputational contribution I_n of most recent interaction n of any given user is estimated in the following way:

$$I_n = I_{b_n} + I_{c_n} = I_{b_n}(1 + \alpha(1 - \frac{1}{A_n + 1})) \tag{1.42}$$

Here, α is the weight of the cumulative part and A_n is the number of sequential activities. If there is no interaction at t_n, this part of interactions has a value of 0. Important property of this component of dynamic reputation is the notion of sequential activities. Two successive interactions made by a user are considered sequential if the time between those two activities is less or equal to the time parameter t_a which represents the time window of interaction. This time window represents maximum time spent by the user to make a meaningful contribution (post a question or answer or leave a comment).

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a} \tag{1.43}$$

If $\Delta_n < 1$ is less than one the number of sequential activities A_n will increase by one, which means that the user is continuing to communicate frequently. On the other hand, large values Δ_n greatly increase the effect of the forgetting factor. This factor plays a major role in updating the total dynamic reputation of a user in each time step (after every recorded interaction):

$$T_n = T_{n-1}\beta^{\Delta_n} + I_n \tag{1.44}$$

Here, β is the forgetting factor. In our implementation of the model, the trust is updated each day for every user irrespective of their activity status. Therefore, the decay itself is a combination of β and Δ_n: the more days pass without recorded interaction from a specific user, the more their reputation decays. Lower values of beta lead to faster decay of trust as shown on figure.

Chapter 2

Driving signals

The complex networks grow through the addition of new nodes, and growing networks models consider that growth is constant over time. This approximation is sufficient for explaining how properties of complex networks can emerge; for example, in the Barabasi-Albert model such as in real systems, we find scaling of degree distribution. Models mostly 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 on daily basis and the users' activity might have bursty nature. We can consider a co-authorship network, where links are created between scientists when they publish a paper. The dynamic of real networks can be complex and highly influenced by non-linear signals. The growth signal; the number of new nodes in each time step; has cycles and trends. Circadian cycles are directly reflected into growth signals and we also find long-range correlations and multifractal properties.

In this chapter, we explain the properties of growth signals, both real and computer-generated. We analyze networks created with a growing network model where the interplay between ageing and preferential attachment shape their structure. We are interested to incorporate non-constant growth signals into the model and measure their impact on the complex networks. Differences between networks with the same number of nodes and links can be observed through connectivity patterns. Figure 2.1 describe used model.

2.1 Growing signals

2.1.1 Long range correlated signals

The main characteristic of long-range correlated time series is power law decay of autocorrelaction function, $C(s) =< x_i x_{i+1} >= s^{-\delta}$. Instead of using correlation function to directly determine type correlations in the signal, in practice is more common to calculate Hurst exponent.

Hurst exponent is used for estimating self-similarity of the time series described with relation $x(ct) = cHx(t)$. Hurst exponent and autocorrelation coefficient δ are connected as $H = 1 - \frac{\delta}{2}$. When $H = 0.5$ signal has short range correlations and is considered to be white noise, while for $H = 1.0$ signal is pink noise. Between this limits $0.5 < H < 1.0$, signal has long range correlations.

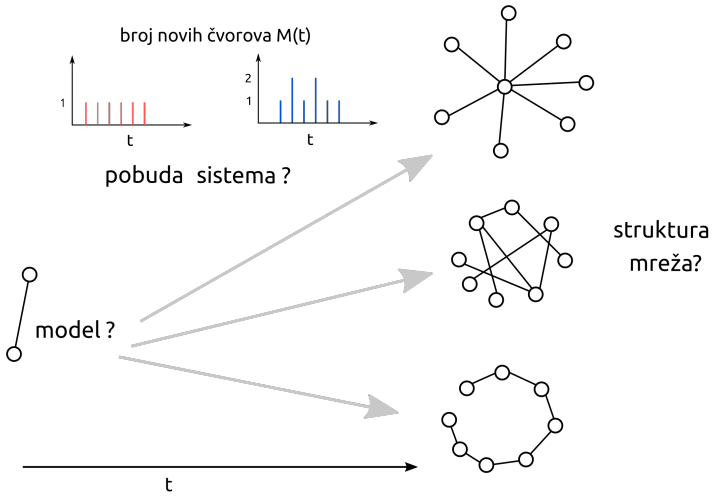

Figure 2.1: Growing network model schema.

Monofractal signals can be generated using Fourier transform method [28]:

- 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
- filter signal $x_q = u_q s$, where s is Fourier transform of autocorrelation function $C(s)$
- the inverse Fourier transform x_i is signal with specific long range correlations

Figure 2.2 shows artificial signals generated using Fourier transform method for different values of Hurst exponents. The obtained signals are round to integers, as in real time series integer values are present. The mean values of signals are close to 4.

For estimation of Hurst exponent from non-stationary signal can be used detrended fluctuation analasys (DFA) [29] [30]. This method removes trends and cycles from the signal, while Hurst exponent is estimated based on residual fluctuations. Signals from real world have usually multifractal structure and can not be described with only one value of Hurst exponent [24]

2.1.2 Multifractal analysis

Multifractal detrended fluctuation analysis (MFDFA) [24, 25] to estimate multifractal Hurst exponent H(q). For given time series $\{x_i\}$ with length N, first we define global profile in the form of cumulative sum, equation 2.1, where where $\langle x \rangle$ represents average of the time series:

$$Y(j) = \sum_{i=0}^{j} (x_i - \langle x \rangle), \quad j = 1, ..., N \tag{2.1}$$

Subtracting the mean of the time series is supposed to eliminate global trends. The profile of the signal Y is divided into $N_s = int(N/s)$ non overlapping segments of length s. If N is not divisible with s the last segment will be shorter. This is handled by doing the same division from the opposite side of time series which gives us $2N_s$ segments. From each segment ν, local trend $p^m_{\nu,s}$ - polynomial

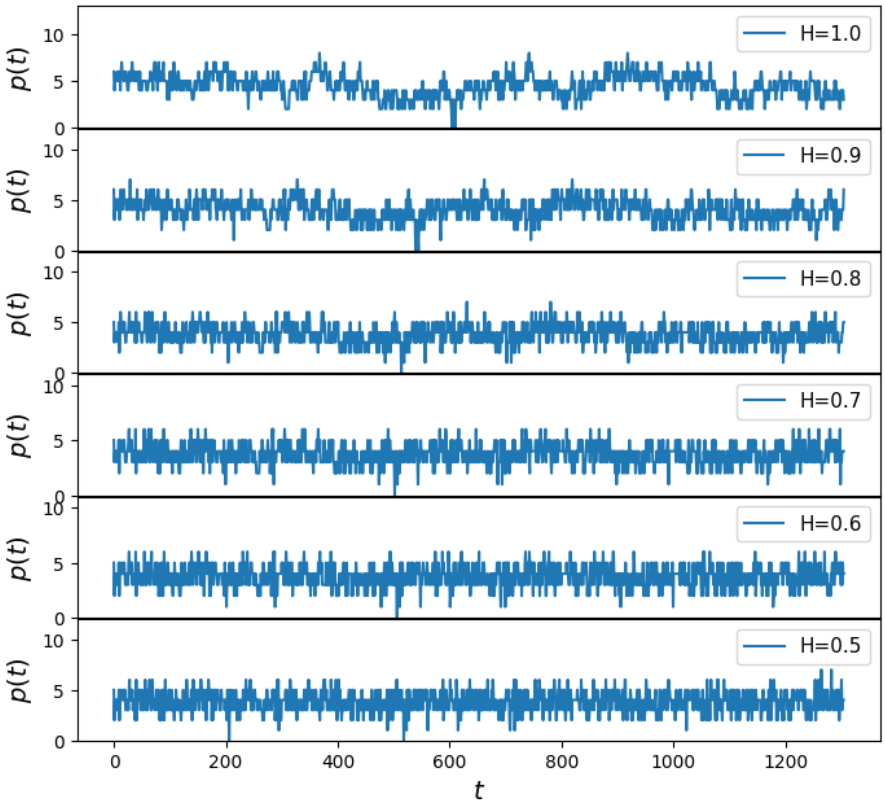

Figure 2.2: Monofractal signals

of order m - should be eliminated, and the variance $F^2(\nu, s)$ of detrended signal is calculated as in equation 2.2:

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^{s} \left[Y(j) - p^m_{\nu,s}(j) \right]^2 \tag{2.2}$$

Then the q-th order fluctuating function is:

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

$$F_0(s) = \exp \left\{ \frac{1}{4N_s} \sum_{\nu}^{2N_s} ln \left[F^2(\nu, s) \right] \right\}, q = 0 \tag{2.3}$$

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 $H(q)$. Multifractal signal has different scaling properties over scales while monofractal is independent of the scale, i.e., H(q) is constant.

2.1.3 Real signals

In this work, we use two different growth signals from real systems figure 1: (a) the data set from TECH community from Meetup social website [36] and (b) two months dataset of MySpace social network [37]. TECH is an event-based community where members organize offline events through the Meetup site [36]. The time unit for TECH is event since links are created only during offline group meetings. The growth signal is the number of people that attend the group's meetings for the

first time. MySpace signal shows the number of new members occurring for the first time in the dataset [37] with a time resolution of one minute. The number of newly added nodes for the TECH signal is N = 3217, and the length of the signal is T = 3162 steps. We have shortened the MySpace signal to T = 20221 time steps to obtain the network with N = 10000 nodes.

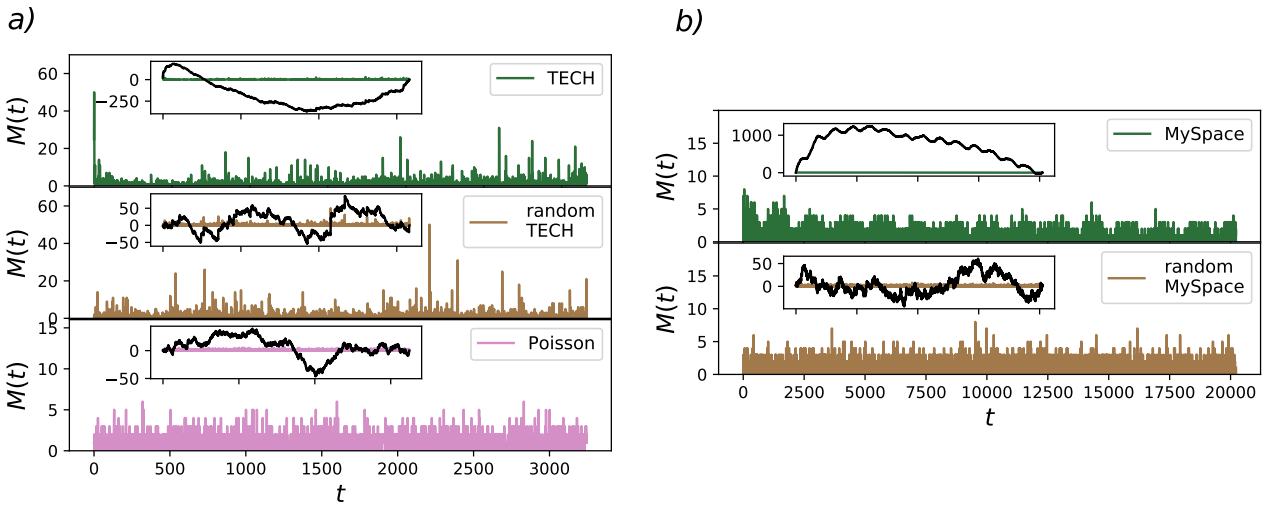

Figure 2.3: Growth signals for TECH (a) and MySpace (b) social groups, their randomized counterparts, and random signal drawn from Poasonian distribution with mean 1. The cumulative signals are shown in insets.

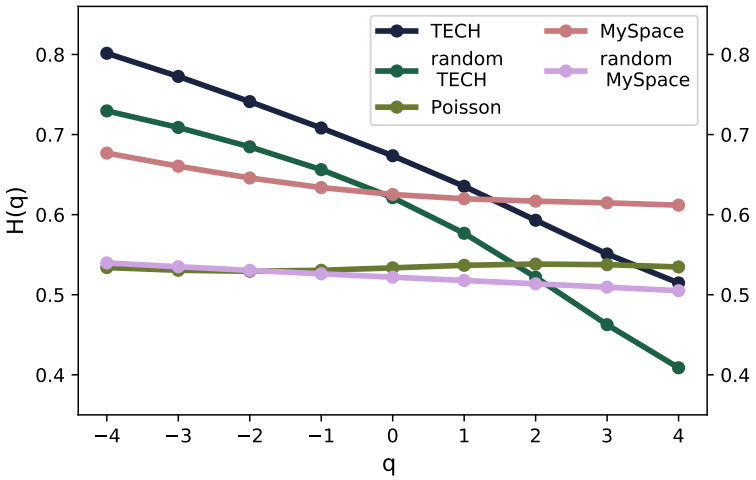

Figure 2.4: Dependence of Hurst exponent on parameter q for all five signals shown in figure 2.3 obtained with MFDFA.

Real growth signals have long-range correlations, trends and cycles [37, 27, 25]. We also generate networks using randomized signals and one computer-generated white- noise signal to explore the influence of these signal's features on the structure of evolving networks. We randomize real signals using reshuffling procedure and keep their length and mean value, the number of added nodes, and probability density function of fluctuations intact, but destroy cycles, trends, and long-range correlations. Besides, we generate a white-noise signal from a Poissonian probability distribution with a mean equal to 1. The length of the signal is T = 3246, and the number of added nodes in the final network is the same as for the TECH signal.

Figures 2.3 (a) and 2.4 show that the TECH signal has long trends and a broad probability density function of fluctuations. The trends are erased from the randomized TECH signal, but the broad distribution of the signal and average value remain intact. MFDFA analysis shows that real signals have long-range correlations with Hurst exponent approximately 0.6 for $q = 2$, figure 2.4. The TECH signal is multifractal, the consequence of both broad probability distribution for the values of time series and different long-range correlations of the intervals with small and large fluctuations. Shuffling of the time series does not destroy the broad distribution of values, the reason for the persistent multifractality of the TECH randomized signal, figure 2.4.

MySpace signal has a long trend with additional cycles that are a consequence of human circadian rhythm, figure 2.3(b). It is multifractal for $q < 0$, and has constant value of $H(q)$ for $q > 0$, figure 2.4. In MFDFA, with negative values of q, we put more emphasis on segments with smaller fluctuations, while for positive q emphasis is more on segments with larger fluctuations [25]. Segments with smaller fluctuations have more persistent long-range correlations in both real signals, see figure 2.4. Randomized MySpace signal and Poissonian signal are monofractal and have short-range with $H = 0.5$ correlations typically for white noise.

2.2 Growing network model with aging nodes

The networks generated with constant growth signal are uncorrelated trees.

To enable formation of clusters in the network new nodes need to create more than one link. We adapt the original model such that at each time step we add $M \geq 1$ new nodes that make $L \geq 1$ links with existing nodes in the network corresponding to probability 1.24. The master equation for N_k, k degree nodes can be written as:

$$\partial_t N_k = \sum_{j=1}^{M(t)} r_{k-j \longrightarrow k} N_{k-j} - \sum_{j=1}^{M(t)} r_{k \longrightarrow k+j} N_k + M(t)\delta_{k,L}. \tag{2.4}$$

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

The equation 2.4 is not solvable in a general case. It was solved for the case $M(t) = 1$ and specific values of parameters α and β using continuous approach [31]. In this work, we use numerical simulations to explore the case when $M(t)$ is a correlated time-varying function and study how these properties influence the structure of generated networks for different values of parameter $-\infty < \alpha \leq -1$ and $\beta \geq 1$ and constant L.

2.3 Structural differences between networks

The advantage this measure has is that it can distinguish between networks generated with the same model parameters. To examine how different growth signals influence the network structure, we use D-measure and compare networks generated with the same model parameters α, β and fixed number

of links per new node L, but different growth signals. The growth of first network is driven by fluctuating signal $M_1 = M(t)$, while the other one grows by constant rate $M_2 =< M(t) >= const$.

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

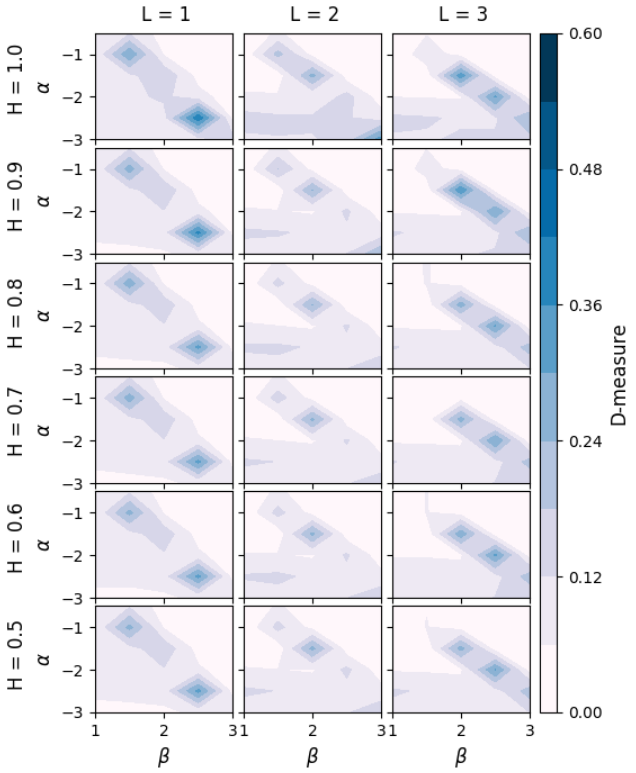

Figure 2.5: D-distance between networks generated with different long-range correlated signals with fixed value of Hurst exponent and networks generated with constant signal M=4.

First, we explore how monofractal signals, see Figure 2.2 shape the structure of complex networks. The D-measure between networks grown with monofractal signal, with $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and constant signal $M = 4$ are shown in figure 2.5. The higher values of D-measure are found in the region of critical line $\beta(\alpha^*)$. The most considerable influence is on networks with scale-free distribution. Comparing D-distance in only one point of phase diagram, for example $L = 1, \alpha = -2.5, \beta = 2.5$, we find correlations in the signal (Hurst exponent is larger), make bigger impact on the network structure. D-measure between networks grown with signal with Hurst exponent $H = 1.0$ and constant signal is $D(H = 1.0, M = 4) = 0.405$, while between networks grown with signal with $H = 0.8$ and constant signal is $D(H = 0.8, M = 4) = 0.316$. For $\alpha > \alpha^*$ networks have similar structural properties and D-measure is close to 0. In the region of networks with stretched exponential degree distribution $\alpha < \alpha^*$ differences are small.

For signals from real communities we find non-zero values of D-measure 2.6. The largest difference between networks is as before along critical line $\beta(\alpha^*)$, for scale free network. For values $\beta < \beta(\alpha^*)$ the structural differences exist, but they become smaller. In the region of gel small world networks $\alpha > \alpha^*$ structural differences are small and close to zero. In the region around critical line

we find that D-measure depends on the properties of the signal. Multifractal signals TECH has the largest impact on network structure; the maximum obtained value of D-measure is $D_{max} = 0.552$. Similar behavior we discover for other multifractal signals, random TECH and MySpace. For networks generated with uncorelated signals, random mySpace and Poisson, difference exists but it is much smaller and comparable with monofractal signals.

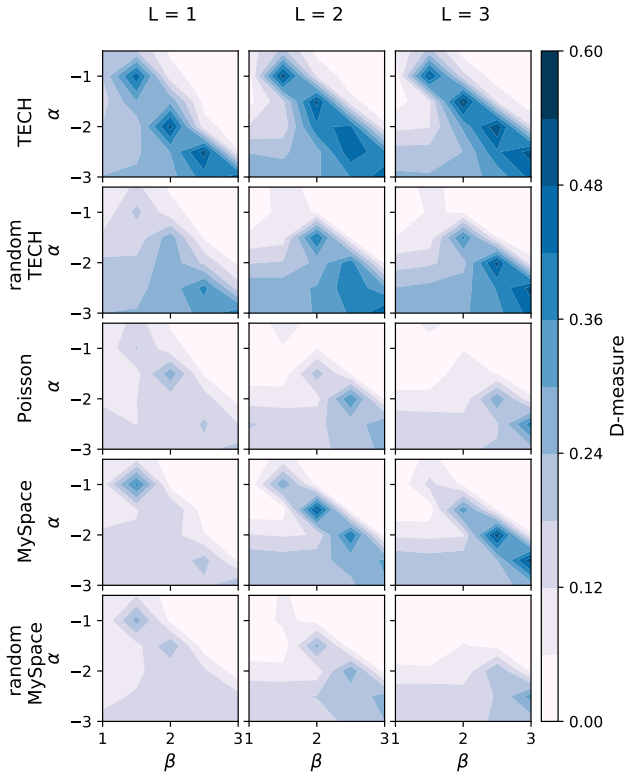

Figure 2.6: The comparison of networks grown with growth signals shown in figure 2.3 versus ones grown with constant signal $M = 1$, for value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in each time step. The compared networks are of the same size.

The position of the critical line slightly moves toward larger β with higher link density L. The addition of more than one node does not influence its position. Although, for fixed network density, we find a critical line independent of the growth signal's properties as can be seen in Figures 2.6, 2.5.

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

2.3.1 The assortativity and clustering

We further explore the assortativity index and clustering coefficient of networks generated with monofractal signals with different values of Hurst exponent. We show results for several ageing model parameters to show the difference between network this model can produce, 2.7. All networks

are disassortative, with a negative degree-degree correlation index. For the values of parameters below critical line, $\alpha = -2.5, \beta = 1.5$ r does not depend on the Hurst exponent. Above the critical line are small-world networks, and they are disassortative with a minimum value of assortativity index $r = -1$, for $L = 1$, indicating the presence of a hub that connects to many nodes. The assortativity index slightly grows with link density.

In the region of critical parameters, the assortativity index depends on the value of the Hurst exponent. The larger influence on the assortativity index have correlated signals, with Hurst exponent $H > 0.8$, so networks become more disassortative, see line for parameters $L = 1, \alpha = -2.5, \beta = 2.5$ in Figure 2.7. The long-range correlations have a stronger effect on the evolution of networks with lower density.

We calculate the mean clustering coefficient, Figure 2.7. For $L = 1$ networks are uncorrelated trees, with clustering coefficient 0. For network density $L > 1$, nodes are organized into clusters. Under the critical line, for parameter $L = 3, \alpha = -2.5, \beta = 1.5$, clustering coefficient is constant and low. Similar values are obtained for clustering coefficient for critical parameters $L = 3, \alpha = -1.5, \beta = 2.0$, but for Hurst exponent $H > 0.8$ clustering coefficient increase. Small world networks, $L = 3, \alpha = -1.5, \beta = 2.5$ are clustered, the value of $< c >$ is high. The value of clustering for networks created with the constant signal is 0.8. Networks grown with white noise signal and signal with H=0.6 have higher values of the clustering, while networks grown with signals that have Hurst exponent larger than 0.6 have the same value of clustering, which is below 0.8.

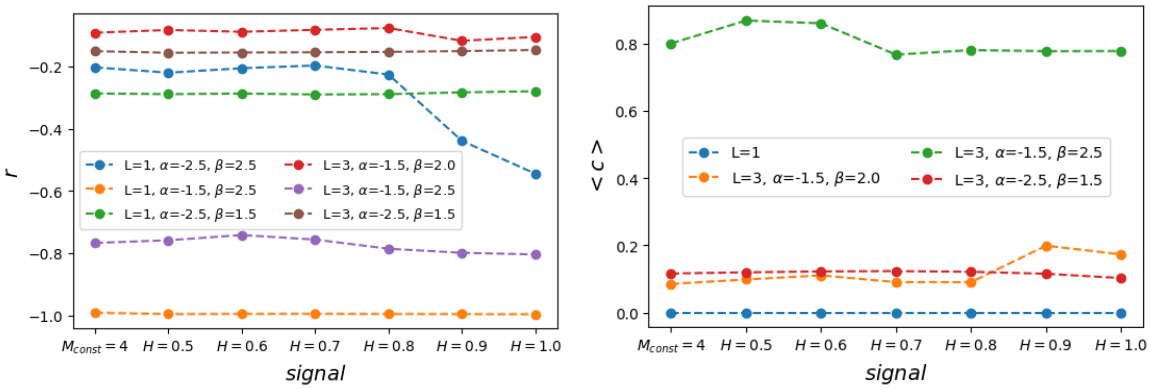

Figure 2.7: Aindex

We examine degree distribution, degree correlations and clustering coefficient of networks generated by real signals, as researchers has shown that these measures provide the sufficient set for discribing structure of complex network. D-measure showed that multifractals have larger influence on networks than monofractals, especially on scale-free networks.

Figure 2.8 shows properties of networks generated with model parameters $L = 2, \alpha = -1.0, \beta = 1.5$, that lies on critical line. The degree distributions $P(k)$ of networks generated with real signals TECH and MySpace have emergence of super-hubs. Degree distributions generated with randomized signals and white noise signal do not differ from degree distribution of networks generated with constant signal. Networks generated with real signals average neighbouring degree $\langle k \rangle_{nn}(k)$ and clustering coefficient $c(k)$ depend on node degree, while in networks generated with constant and randomized signals they weakly depend on the degree k.

We also find structural differences between networks, obtained with model parameters under the critical line $\alpha < \alpha^*$, see Figure 2.9. The difference is mostly found for TECH signal. Degree distribution $P(k)$ shows emergence of hubs in networks grown with TECH signal, while the randomized and Poisson signal are more similar to networks grown with constant signal. MySpace signal; whose

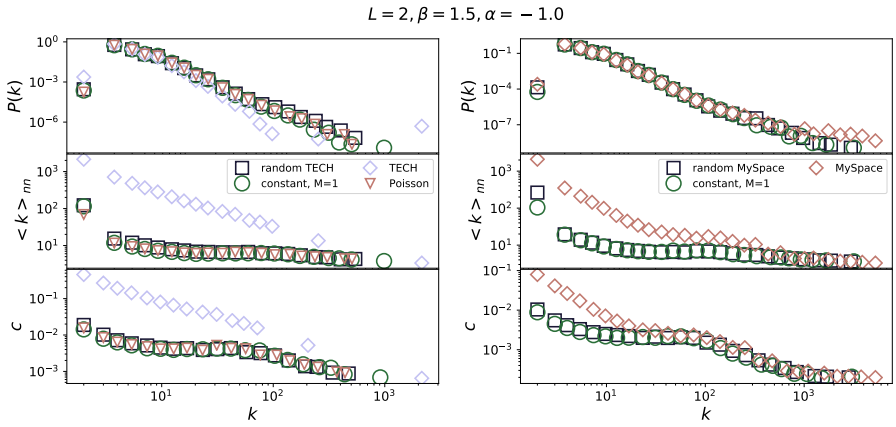

Figure 2.8: Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $\alpha = -1.0$, $\beta = 1.5$ and $L = 2$ for all networks. The networks are from scale-free class.

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.

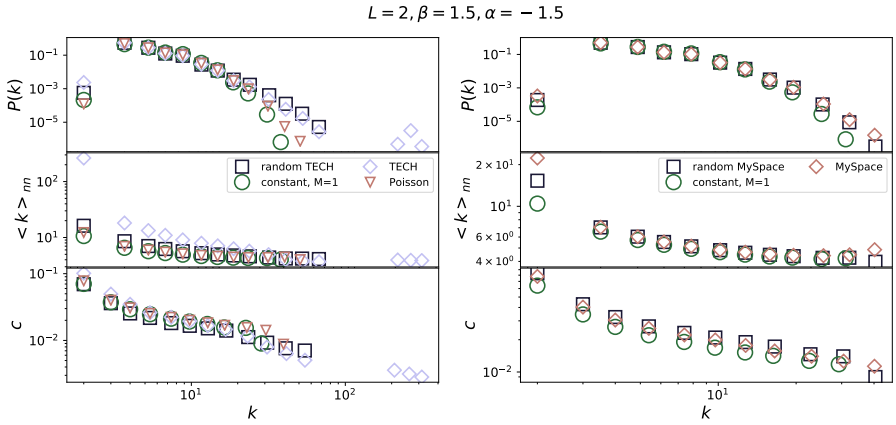

Figure 2.9: Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $L = 2, \alpha = -1.5$, $\beta = 1.5$. The networks have stretched exponential degree distribution.

The properties of time-varying signal do not influence the topological properties of small-world gel networks, Figure 2.10. Here model promote existence of hubs. As this is mechanism through which the fluctuations alter the structure of evolving networks, the properties of the signal are not relevant.

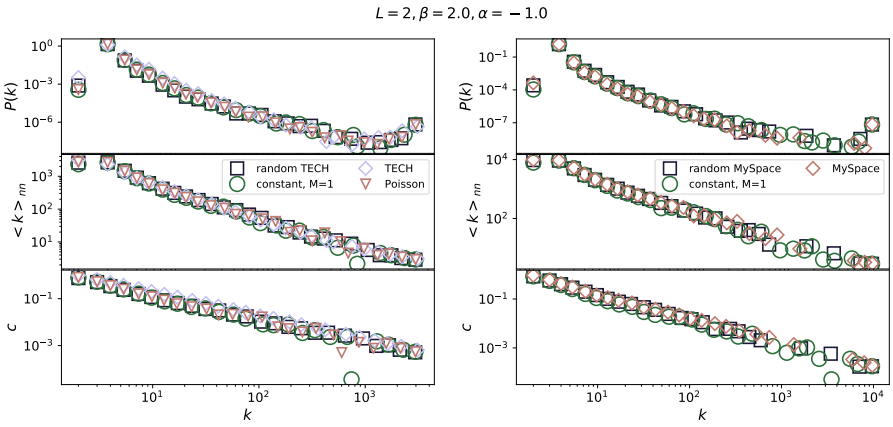

Figure 2.10: Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $L = 2, \alpha = -1.0, \beta = 2.0$. Generated networks have scale-free properties.

2.4 Conclusions

We demonstrate that the resulting networks' structure depends on the features of the time-varying signal that drives their growth. The previous research [32, 33] indicated the possible influence of temporal fluctuations on network properties. Our results show that the temporal properties of growth signals generate networks with power-law degree distribution, non-trivial degree-degree correlations, and clustering coefficient even though the local linking rules, combined with constant growth, produce uncorrelated networks for the same values of model parameters [20].

We observe the most substantial dissimilarity in network structure along the critical line, the values of model parameters for which we generate networks with broad degree distribution. Figure 2.6 shows that dissimilarity between networks grown with time-varying signals and ones grown with constant signals always exists along this line regardless of the features of growth signal. However, the magnitude of this dissimilarity strongly depends on these features. We observe the largest structural difference between networks grown with multifractal TECH signal and networks that evolve by adding one node in each time step. The identified value of D-measure is similar to one calculated in the comparison between sub-critical and super-critical Erdös–Rényi graphs [16] indicating the considerable structural difference between these networks. Our findings are further confirmed in figure **??**(b). The networks generated with signals that have trends and long-range temporal correlations differ the most from those grown with the constant signal. Our results show that even white-noise type signals can generate networks significantly different from ones created with constant signal for low values of α^*.

The value of D-measure declines fast as we move away from the critical line, figure 2.6. The main mechanism through which the fluctuations influence the structure of evolved networks is the emergence of hubs and super hubs. For values of $\alpha << \alpha^*$, the nodes attache to their immediate predecessors creating regular networks without hubs. For $\alpha \sim \alpha^*$ graphs have stretched exponential degree distribution with low potential for the emergence of hubs. Still, multifractal signal TECH enables the emergence of hub even for the values of parameters for which we observe networks with stretched-exponential degree distribution in the case of constant growth figure **??**(a). By definition, small-world gels generated for $\alpha > \alpha^*$ have super-hubs [20] regardless of the growth signal, and therefore the effects that fluctuations produce in the growth of networks do not come to the fore for values of model parameters in this region of $\alpha - \beta$ plane.

Evolving network models are an essential tool for understanding the evolution of social, biological, and technological networks and mechanisms that drive it [**?**]. The most common assumption is that these networks evolve by adding a fixed number of nodes in each time step [**?**]. So far, the focus on developing growing network models was on linking rules and how different rules lead to networks of various structural properties [**?**]. Growth signals of real systems are not constant [33, 32]. They are multifractal, characterised with long-range correlations [33], trends and cycles [34]. Research on temporal networks has shown that temporal properties of edge activation in networks and their properties can affect the dynamics of the complex system [8]. Our results imply that modeling of social and technological networks should also include non-constant growth and that its combination with local linking rules can significantly alter the structure of generated networks.

Chapter 3

Groups growth model

3.1 Introduction

Social groups, informal or formal, are mesoscopic building elements of every socio-economic system that direct its emergence, evolution, and disappearance []. The examples span from countries, economies and science to society in general. Settlements, villages, towns and cities are formal and highly structured social groups of countries. Their organisation and growth determine the functioning and sustainability of every society [35]. Companies are the building blocks of an economic system and their dynamics are important indicators of the level of its development [36]. Scientific conferences, as scientific groups, enable fast dissemination of the latest results, exchange and evaluation of ideas as well as a knowledge extension, and thus are an integral part of science [37]. The membership of individuals in various social groups, online and offline, can be essential when it comes to the quality of their life [38, 39, 40]. Therefore, it is not surprising that the social group emergence and evolution are at the center of the attention of many researchers [41, 42, 43, 44].

Along with massive data sets comes the need to develop methods and tools for their analysis and modeling. Methods and paradigms from statistical physics have proven to be very useful in studying the structure and dynamics of social systems [45]. The main argument for using statistical physics to study social systems is that they consist of many interacting individuals. Due to this, they exhibit different patterns in their structure and dynamics, commonly known as *collective behavior*. While building units of a social systems can be characterized by many different properties, only few of them enforce collective behavior in the systems. The phenomenon is known as *universality* in physics and is commonly observed in social systems such as in voting behavior [46], or scientific citations [47]. It indicates the existence of the universal mechanisms that govern the dynamics of the system [].

The availability of large-scale and long-term data on various online social groups has enabled the detailed empirical study of their dynamics. The focus was mainly on the individual groups and how structural features of social interaction influence whether individuals will join the group [48] and remain its active members [37, 49]. The study on LiveJournal [48] groups has shown that decision of an individual to join a social group is greatly influenced by the number of her friends in the group and the structure of their interactions. The conference attendance of scientists is mainly influenced by their connections with other scientists and their sense of belonging [37]. The sense of belonging

of an individual in social groups is achieved through two main mechanisms [49]: expanding of the social circle at the beginning of joining the group and strengthening of the existing connections in the later phase. The dynamics of social groups depend on their size []. Analysis of the evolution of large-scale social networks has shown that edge locality plays a critical role in the evolution of social networks [50]. Small groups are more cohesive with continued membership, while large groups tend to change their active members constantly [**?**]. These findings help us understand the growth of a single group, the evolution of its social network, and the influence of the network structure on the group growth. However, how the growth mechanisms influence the distribution of members of one social system among groups is still anecdotal.

Furthermore, it is not clear whether the growth mechanisms of social groups are universal or system-specific. The size distribution of social groups has not been extensively studied. Rare empirical evidence of the size distribution of social groups indicates that it follows power-law behavior [51]. However, the distribution of company sizes follows log-normal behavior and remains stable over decades [52, 53]. Analysis of the sizes of the cities shows that the distribution of all cities also follows a log-normal distribution. In contrast, the distribution of the largest cities resembles Zipf's distribution [54].

A related question that should be addressed is whether we can create a unique yet relatively simple microscopic model that reproduces the distribution of members between groups and explains the differences observed between social systems. French economist Gibrat proposed a simple growth model to reproduce the observed log-normal size distribution of companies and cities. However, the analysis of the growth rate of the companies [52] has shown that growth mechanisms are different from ones assumed by Gibrat. In addition, the analysis of the growth of three online social networks showed that population growth is not determined by the population size and spatial factors, and it deviates from Gibrat's law [55]. Other mechanisms, for instance, growth through diffusion, have been used for modeling and prediction of rapid group growth [56]. However, the growth mechanisms of various social groups and the source of the scaling observed in socio-economic systems remain hidden.

Here we analyze the size distribution of formal social groups in two different systems: Meetup online platform and subreddits on Reddit. We are interested in the scaling behavior of size distributions and the distribution of growth rates. Empirical analysis of the dependence of growth rates, shown in this work, indicates that growth cannot be explained through Gibrat's model. Here we contribute with a simple microscopic model that incorporates some of the findings of previous research [48, 51]. We show that the model can reproduce size distributions and growth rate distributions for both studied systems. Moreover, the model is flexible and can produce a broad set of size distributions depending on the value of model parameters.

The paper is organized as follows: in Section 3.2 we describe the data, while in Section 3.3 we present our empirical results. In Section 3.4 we introduce model parameter and rules. In section 3.5 we demonstrate that model can reproduce the growth of social groups in both systems and show the results for different values of model parameters. Finally, in Section 3.6, we present concluding remarks and discuss our results.

3.2 Data

We analyse the growth of social groups from two widely used online platforms: Reddit and Meetup. Reddit [1] enables sharing diverse web content and members of this platform interact exclusively online through posts and comments. The Meetup [2] allows people to use online tools to organize offline meetings. The building elements of the Meetup system are topic-focused groups, such as food lovers or ICT and data science professionals. Due to their specific activity patterns - events where members meet face-to-face - Meetup groups are geographically localised.

We compiled the Reddit data from https://pushshift.io/. This site collects data daily and, for each month, publishes merged comments and submissions in the form of JSON files. Specifically, we focus on subreddits - social groups of Reddit members interested in a specific topic. We select subreddits active in 2017 and follow their growth from their beginning until $2011 - 12$. The considered dataset contains 17073 subreddits with 2195677 active members, with the oldest originating from 2006 and the youngest being from 2011. For each post under a subreddit, we extracted the information about the member-id of the post owner, subreddit-id, and timestamp. As we are interested in the subreddits growth in the number of members, for each subreddit and member-id we selected timestamp when member made a post for the first time. Finally, in the dataset we include only subreddits active at least two months.

The Meetup data were downloaded in 2018 using public API. The Meetup platform was launched in 2003, and at the moment we accessed the data, there were more than 240 000 active groups. For each group, we extracted information about the date it had been founded, its location, and the total number of members. We focused on the groups founded from 2003 until 2017 in big cities London and New York, where Meetup platform achieved considerable popularity. We considered groups active at least two months. There were 4673 groups with 831685 members in London and 4752 groups with 1059632 members in New York. In addition, we extracted the id of each member in the group and the information about organised events. This allowed us to obtain the date when a member joined a group, which is the first time she attended group event.

In both systems, we approximated the timestamp when the member joined the group. Based on this information, we can calculate the number of new members per month $N_i(t)$, the group size $S_i(t)$ at each time step, and the growth rate for the group in both systems. The time step for both systems is one month. The size of the group i at time step t is the number of members that joined that group ending with the month, i.e., $S_i(t) = \sum_{k=t_{i0}}^{k=t} N_i(t)$, where t_{i0} is the time step in which the group i was created. We do not consider when a member leaves a group or subreddit since this kind of information is not available to us. For these reasons, the size of considered groups is a non-decreasing function. The growth rate $R_i(t)$ at step i is obtained as logarithm of successive sizes $R = log(S_i(t)/S_i(t-1))$.

While the forms of communication between members and activities that members engage in differ in those two systems, some common properties exist between them. Members can form a new groups and join existing ones in both systems. Furthermore, each member can belong to an unlimited number of groups. For these reasons, we can use the same methods to study and compare the formation of groups in Reddit and Meetup.

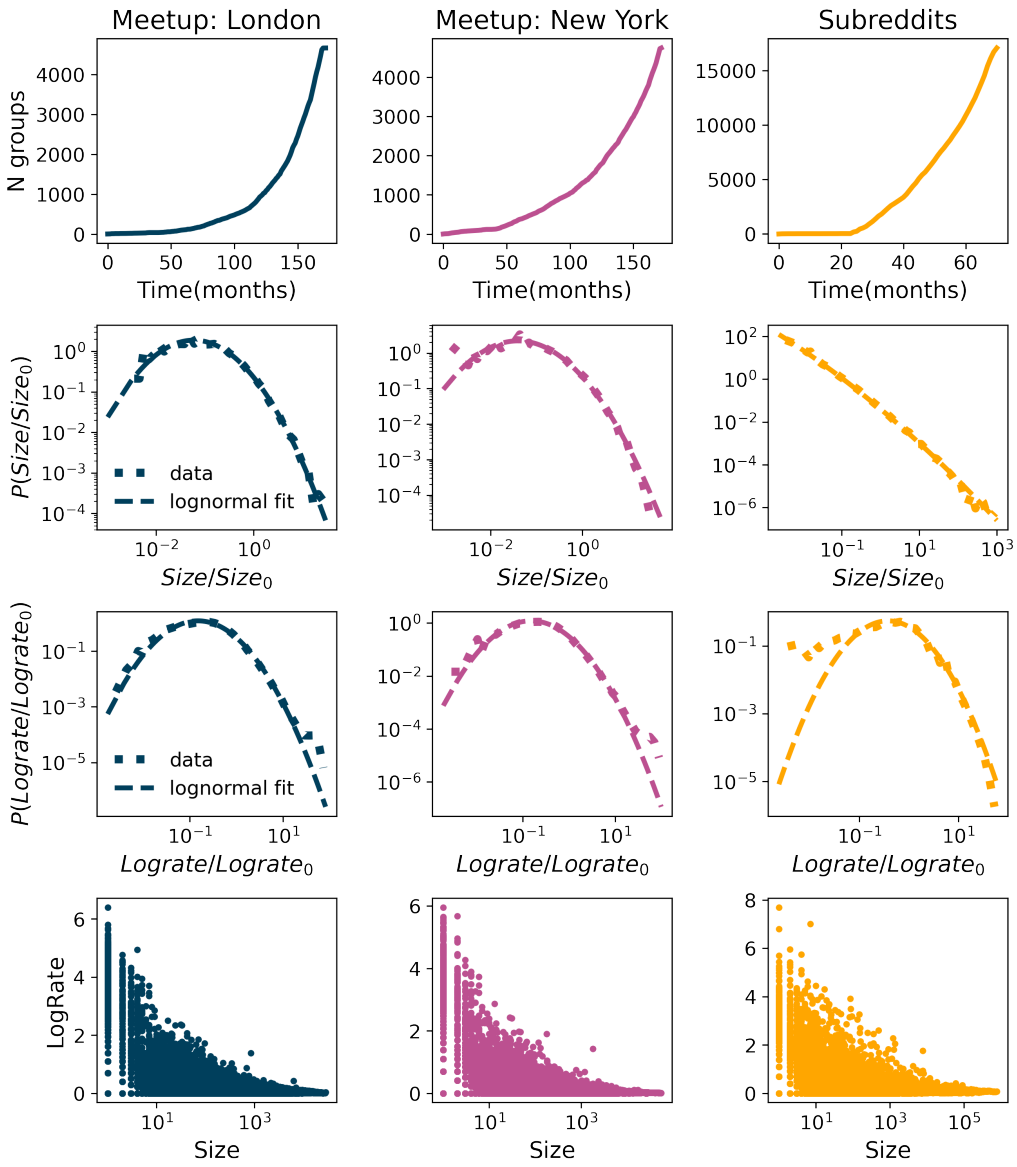

Figure 3.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011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

3.3 Empirical analysis of social group growth

Figure 3.1 summarize properties of the groups in Meetup and Reddit systems. The number of groups grows exponentially over time. Nevertheless, we notice that Reddit has substantially larger number of groups than Meetup. The Reddit groups are prone to engage more members in a shorter period of time. Size of the Meetup groups is in the range from several members up to several tens of thousands of members, while sizes of subreddits are between a few tens of members up to several millions. The distributions of group sizes follows the lognormal distribution

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

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

where S is the group size and μ and σ are parameters of the distribution. We used package [57] to fit Eq. 3.1 to Reddit and Meetup data and found that distribution of groups sizes for Meetup groups in London and New York follow similar distributions with the values of parameters $\mu = -0.93$, $\sigma = 1.38$ and $\mu = -0.99$ and $\sigma = 1.49$ for London and New York respectively. The distribution of sizes of subreddits also has the log-normal shape with parameters $\mu = -5.41$ and $\sigma = 3.07$. Even though these distributions are from the same class, for subreddits we find broader distribution that may resemble power-law distribution. Our analysis shown in Supportive Information (SI) confirms that the distribution exhibits a log-normal behavior, see SI-Table 1 and SI-Fig. 1.

The log-normal distributions can be generated by multiplicative processes [22]. If there is a quantity with size $S_i(t)$ at time step t, it will grow so after time period δ the size of the quantity is $S(t + \Delta t) = S(t)r$, where r represents a random process. The Gibrat law states that growth rates r are uncorrelated and do not depend on the current size. In order to describe the growth of social groups, we calculate the logarithmic growth rates defined as $R = log\frac{S_t}{S_{t-\Delta t}}$. According to Gibrat law, the distribution of sizes follow log-normal distribution. For logarithmic growth rates expected distribution is normal, or as it is shown in many studies it is better explained with Laplacian ("tent shaped") distribution [58], [59]. In figure 3.1 we calculate distributions of log-rates. For both systems, log-rates are very well approximated with log-normal distribution. The Fig. 3.1 shows that log-rates depend on the groups size, especially for the smaller and medium size groups. Our empirical analysis implies that the growth of Meetup and Reddit groups violates the basic assumptions of the Gibrat's law [60, 61], and thus, this growth can not be explained as a simple multiplicative process.

We are considering a relatively large time period for online groups. The fast expansion of Information Communications Technology (ICT) led to change of how members access online systems. With the use of smartphones the online systems became more available, which led to exponential growth of ICTs systems, figure 3.1 and potential change in the mechanisms that influence growth of social groups in them. For these reasons we aggregate groups according to year they were founded for each of the three data sets and look at the distributions of these sizes in the year 2017 for Meetup groups and 2011 for Reddit. For each year and each of the three data sets we calculate the average size of the groups that were created in a year $y < S^y >$. We normalize the size of the groups created in year y with corresponding average size $s_i^y = S_i^y / < S^y >$ and calculate the distribution of the normalized sizes for each year. The distribution of normalized sizes for all years and all data sets is shown in figure 3.2. All distributions exhibit log-normal behavior. Furthermore, the distributions for the same data set and different years follow a universal curve with same value of parameters μ and σ. The universal behavior is observed for distribution of normalized log-rates as well, see Fig. 3.2 (bottom panel). These results indicate that growth of the social groups did not change due to increased growth of members in systems. Furthermore, it implies that the growth is independent of the size of the whole data set.

3.4 Model

Growth of social groups can not be explained with the simple rules of Gibrat's law. Previous research on group growth and longevity has shown that social connections with members of a group influence individual's choice to join that group [56, 51]. Moreover, individual's interests and the need to discover new content or activity also influence diffusion of individuals between groups. Furthermore, social systems constantly grow since new members join every minute. The properties of the growth signal that describes the arrival of new members influence both dynamics of the system [62, 63] and the structure of social interactions [64]. Furthermore, number of social groups in the social systems is not constant. They are constantly created and destroyed.

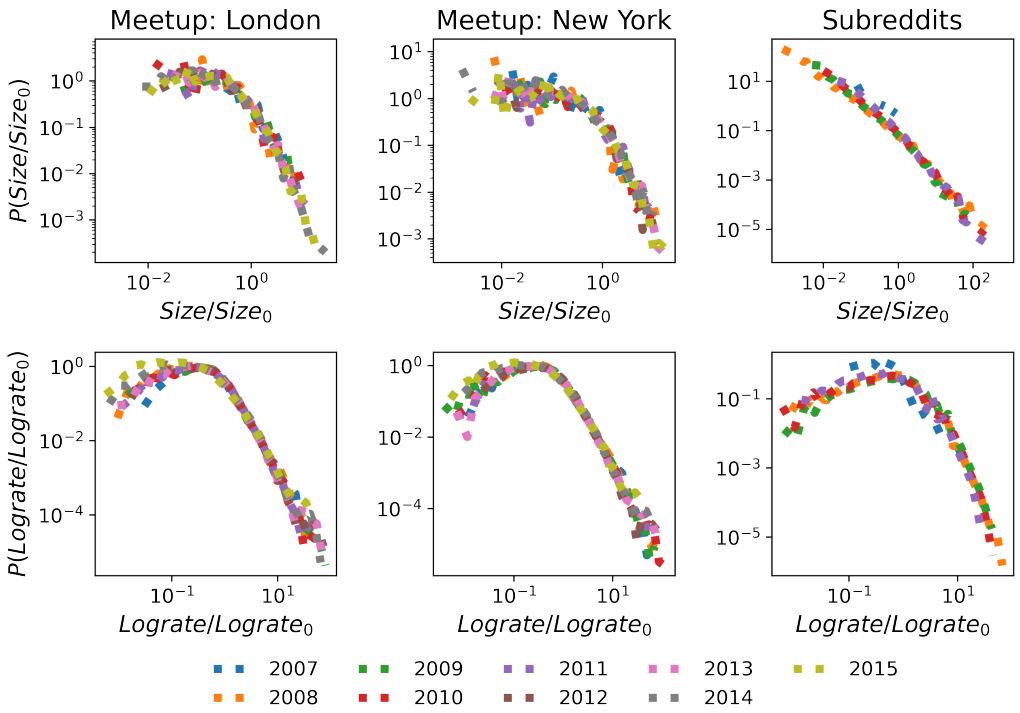

Figure 3.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

In Ref. [51] authors propose the co-evolution model of the growth of social networks. In this model, authors assume that social system evolves trough co-evolution of two networks: network of social contacts between members and network of members' affiliations with groups. This model addresses the problem of growth of social networks that includes both linking between members and social group formation. In this model, a member of a social system selects to join a group either through random selection or according to her social contacts. In the case of random selection, there is a selection preference toward larger groups. If member chooses to select a group according to her social contacts, the group is selected randomly from the list of groups with which her friends are already affiliated.

While the co-evolution model [51] was not created with the intent of studying the growth and size distribution of social groups, authors show that their model is able to reproduce distribution of group sizes for several online social networks that follow power-law distribution. Our empirical analysis, shown in Sec. 3.3 shows that distribution of group sizes is not always power-law, indicating that certain mechanisms proposed in co-evolution model are not universal for all social systems. To fill the gap in understanding how social groups in social system grow, we propose a model of group growth that combines random and social diffusion between groups but following different rules than co-evolution model [51].

Figure 3.3 shows a schematic representation of our model. Similar to co-evolution model [51], we represent social system with two evolving networks, see Fig. 3.3. One network is bipartite network which describes the affiliation of individuals to social groups $\mathcal{B}(V_U, V_G, E_{UG})$. This network consists of two partitions, members V_U and groups V_G, and set of links E_{UG}, where a link $e(u, g)$ between a member u and a group g represents the member's affiliation with that group. Bipartite network grows through three activities: arrival of new members, creation of new groups, and through members joining groups. By definition, in bipartite networks links only exist between

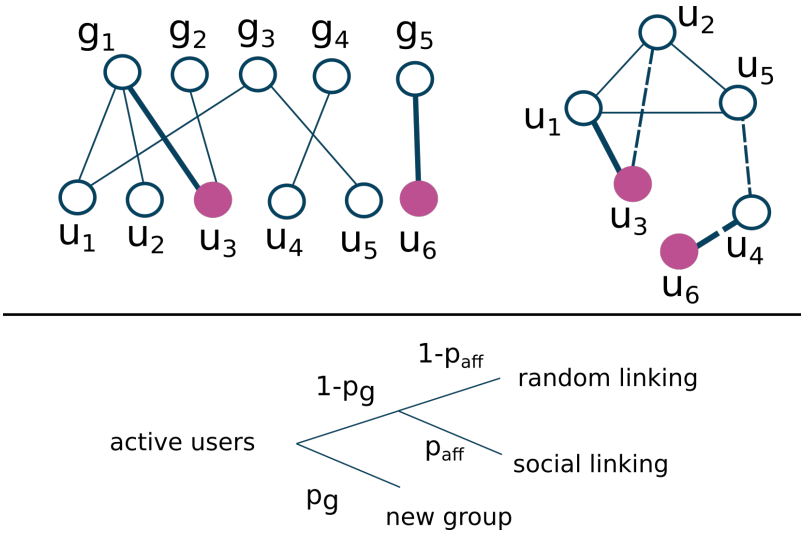

Figure 3.3: The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model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

nodes belonging to different partitions. However, as we explained above, social connections affect whether a member will join a certain group or not. In the simplest case, we could assume that all members belonging to a group are connected with each other. However, previous research on this subject [49, 48, 51] has shown that the existing social connections of members in a social group are only a subset of all possible connections. For these reasons, we introduce another network $\mathcal{G}(V_U, E_{UU})$ that describes social connections between members. The social network grows through addition of new members to the set V_U and creation of new links between them. The member partition in bipartite network $\mathcal{B}(V_U, V_G, E_{UG})$ and set of nodes in members' network $\mathcal{G}(V_U, E_{UU})$ are identical.

For convenience, we represent bipartite and member networks with adjacency matrices B and A. The element of matrix B_{ug} equals one if member u is affiliated with group g, and zero otherwise. In matrix A, the element $A_{u_1u_2}$ equals one if members u_1 and u_2 are connected and zero otherwise. The neighbourhood of member u $\mathcal{N}_u$ is a set off groups that member is affiliated with. On the other hand, the neighbourhood of group g $\mathcal{N}_g$ is a set of members affiliated to that group. The size of set $\mathcal{N}_g$ equals to the size of the group g S_g.

In our model, the time is discrete and networks evolve through several simple rules. In each time step we add $N_U(t)$ new members and increase the size of the set V_U. For each newly added member we create the link to a randomly chosen old member in the social network G. This condition allows each member to perform social diffusion [56], i.e., to choose a group according to her social contacts. Not all members from set V_U are active in each time step. Only a subset of existing members is active in one time step. Activity of old members is a stochastic process and is determined by parameter p_a; every old member is activated with probability p_a. Old members activated in this way and new members make a set of active members $\mathcal{A}_U$ at time t.

The group partition V_G grows through creation of new groups. Each active member $u \in \mathcal{A}_U$ can decide with probability p_g to create a new group, or to join an already existing one with probability $1 - p_g$.

If the active member u decides that she will join an existing group, she first needs to a choice of this group. A member u with probability p_{aff} decides to select a group based on her social connections. For each active member, we look at how many social contacts she has in each group. The number of social contacts s_{ug} that member u has in group g equals to the overlap of members affiliated with a group g and social contacts of member u, and is calculated according to

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

Member u selects an old group g to join according to probability P_{ug} that is proportional to s_{ug}. Member only considers groups with which it has no affiliation. However, if an active member decides to neglect her social contacts in the choice of the social group, she will, with probability $1 - p_{aff}$, select a random group from the set V_G with which she is not yet affiliated.

After selecting the group g, a member joins that group and we create a link in bipartite networks between a member u and a group g. At the same time, member selects X members of a group g which do not belong to her social circle and creates social connections with them. As a consequence of this action, we create X new links in network $\mathcal{G}$ between member u and X members from group g.

The evolution of bipartite and social networks, and consequently growth of social groups, is determined by parameters p_a, p_g and p_{aff}. Parameter p_a determines the activity level of members and takes values between 0 and 1. Higher values of p_a result in higher number of active members and thus faster growth of number of links in both networks, as well as the size and number of groups. Parameter p_g in combination with parameter p_a determines the growth of the set V_G. $p_g = 1$ means that members only create new groups, and the existing network consists of star-like subgraphs with members being a central nodes and groups as leafs. On the other hand $p_g = 0$ means that there is no creation of new groups and the bipartite network only grows through addition of new members and creation of new links between members and groups.
Parameter p_{aff} is especially important. It determines the importance of social diffusion. $p_{aff} = 0$ means that social connections are irrelevant and the choice of group is random. On the other hand, $p_{aff} = 1$ means that only social contacts become important for group selection.
Our model is different from co-evolution model Ref. [51]. In our model p_{aff} is constant and the same for all members. In the co-evolution model this probability depends on members degree. The members are activated in our model with probability p_a, while in co-evolution model members are constantly active from the moment they are added to a set V_U until they become inactive after time t_a. Time t_a differs for every member and is drawn from exponential distribution with rate λ. In co-evolution model the number of social contacts that member has within the group is irrelevant for the group selection. On the other hand, in our model members tend to choose more often groups in which there is a greater number of their social contacts. While in our model, in the case of random selection of a group, member selects a uniformly at random a group that she is not affiliated with, in the co-evolution model the choice of group is preferential.

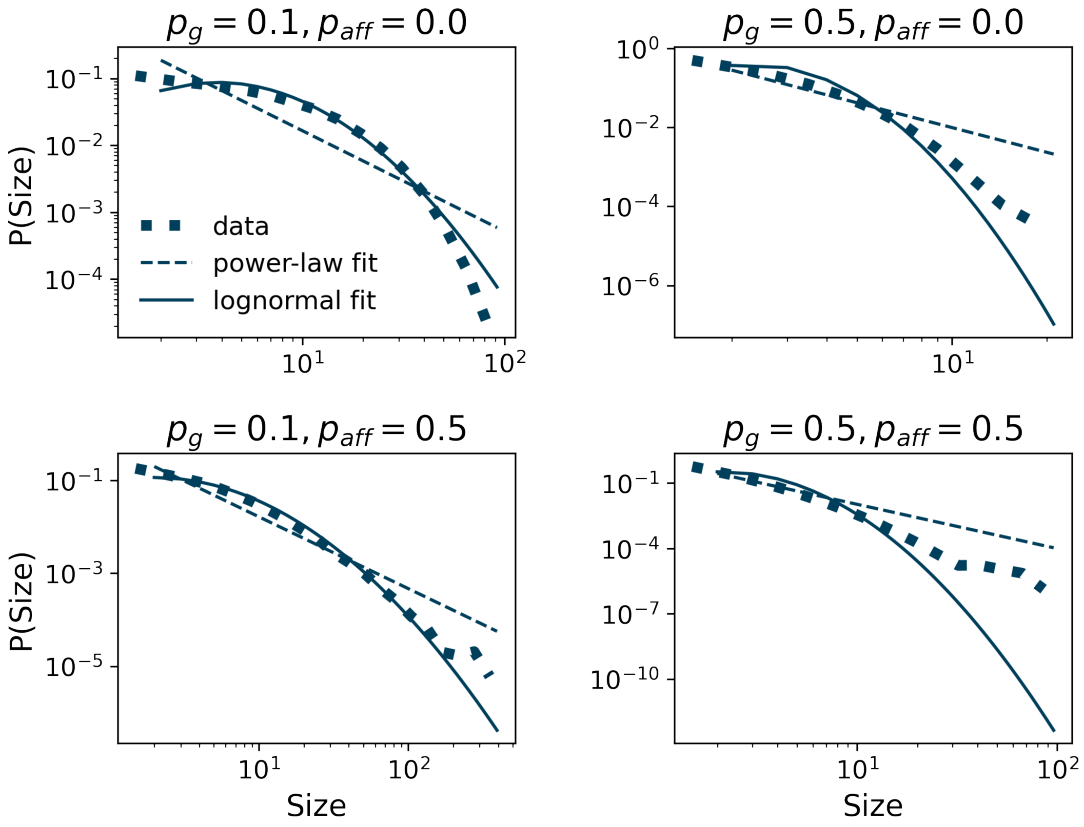

Figure 3.4: The distribution of sizes for constant growth of members, $N = 30$. The probability that members are active is fixed to $p_a = 0.1$, while we vary the probability for the creation of groups p_g and affiliation linking p_{aff}

3.5 Results

The differences between our and co-evolution model, described in previous sections, at first glance may appear small. However, they lead to huge differences in the distribution of the size of social groups. The distribution of group sizes in co-evolution model is a power-law. Our model adds flexibility to produce groups with log-normal size distribution. This expands classes of social systems that can be modeled.

3.5.1 Model description

First, we explore the properties of size distribution depending on parameters p_g and p_{aff}, and fixed value of activity parameter p_a and constant number of members added in each step $N(t) = 30$. The parameter X is set to value 25 for all simulations presented in this work. Our detailed analysis of the results for different values of parameter X shows that these results are independent of the value of parameter X.

Figure 3.4 shows some of the selected results and their comparison with power-law and log-normal fits. We see that values of both p_g and p_{aff} parameters, influence the type and properties of size distribution. For low values of parameter p_g, left column in Fig. 3.4, the obtained distribution is log-normal. The width of the distribution depends on p_{aff}. Higher values of p_{aff} lead to a broader distribution.
As we increase p_g, right column Fig. 3.4, the size distribution begins to deviate from log-normal

distribution. The higher the value of parameter p_g, the faster grows the number of groups available to members. For the value of parameter $p_g = 0.5$, every second active member creates a group in each time step, and the number of groups increases fast. How members are distributed in these groups depends on the value of parameter p_{aff}. When $p_{aff} = 0$, social connections are irrelevant for the choice of the group and members choose groups at random. The obtained distribution slightly deviates from log-normal, especially for large group sizes. In this case large groups sizes become more probable than in the case of log-normal distribution. The non zero value of parameter p_{aff} means that the choice of group becomes dependent on social connections. When member chooses a group according to her social connections, larger groups have higher probability to be affiliated with social connections of active members, and thus this choice resembles preferential attachment. For these reasons, the obtained size distribution has more broad tail than log-normal distribution, and begins to resemble power-law distribution.

3.5.2 Modeling real systems

The social systems do not grow at constant rate. In Ref. [64] authors have shown that features of growth signal influence the structure of social networks. For these reasons we use the real growth signal from Meetup groups located in London and New York, and Reddit community to simulate the growth of the social groups in these systems. Figure 3.5 top panel shows the time series of the number of new members that join each of the three systems each month. All three systems have relatively low growth at the beginning, and than the growth accelerates as the system becomes more popular.

We also use empirical data to estimate p_a, p_g and p_{aff}. Probabilities that old members are active p_a and that new groups are created p_g can be approximated directly from the data. Activity parameter p_a is the ratio between the number of old members that were active in month t and the total number of members in the system at time t. Figure 3.5 middle row shows the variation of parameter p_a during the considered time interval for each system. The values of this parameter fluctuates between 0 and 0.2 for London and New York based Meetup groups, while its value is between 0 and 0.15 for Reddit. To simplify our simulations we assume that p_a is constant in time, and estimate its value as its median value during the 170 months for Meetup systems, and 80 months of Reddit system. For Meetup groups based in London and New York $p_a = 0.05$, while Reddit members are more active on average and $p_a = 0.11$ for this system.
Figure 3.5 bottom row shows the evolution of parameter p_g for the three considered systems. The p_g in month t is estimated as the ratio between the groups created in month t $Ng_{new}(t)$ and the total number of groups that month $Ng_{new}(t) + Ng_{old}(t)$, i.e., $p_g(t) = \frac{Ng_{new}(t)}{N_{new}(t)+N_{old}(t)}$. We see from Fig. 3.5 that $p_g(t)$ has relatively high values at the beginning of the system's existence. This is not surprising. At the beginning these systems have relatively small number of groups and often cannot meet the needs for content of all their members. As the time passes, the number of groups grows, as well as content offerings within the system, and members no longer have a high need to create new groups. Figure 3.5 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} is not possible to estimate directly from the empirical data. For these reasons, we simulate the growth of social groups each of the three systems with the time series of new members obtained from the real data and estimated values of parameters p_a and p_g, while we vary the value of p_{aff}. For each of the three systems, we compare the distribution of group sizes obtained from simulations for different values of p_{aff} with ones obtained from empirical analysis using Jensen

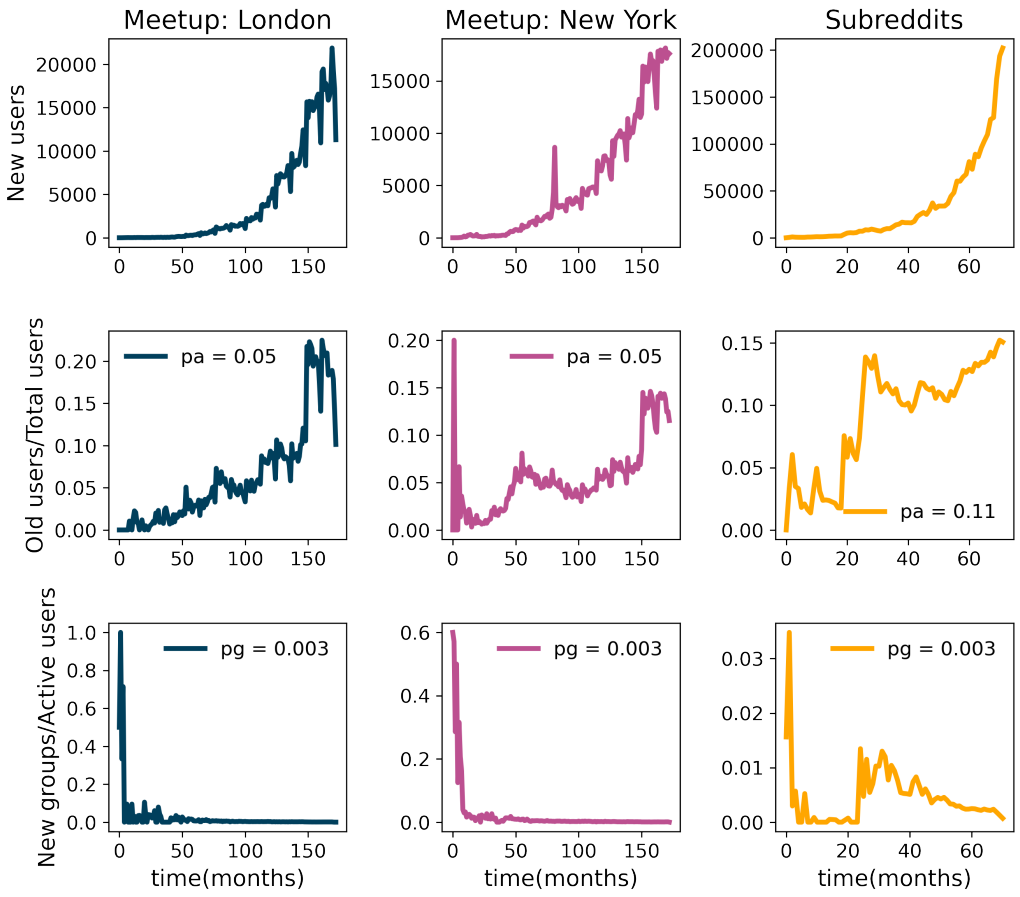

Figure 3.5: The time series of number of new members (top panel), ratio between old members and total members in the system (middle panel), and ratio between new groups and active members(bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

Shannon (JS) divergence. The JS divergence [**?**] between two distributions P and Q is defined as

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

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

Figure 3.6 shows the comparison between the empirical and simulation distribution of group sizes for three considered systems. We see that empirical distributions for Meetup groups based in London and New York are perfectly reproduced by the model and chosen values of parameters. In the case of Reddit, the distribution is very broad, and the tail of distribution is well reproduced by the model. The bottom row of Fig. 3.6 shows the distribution of logarithmic values of growth rates of groups obtained from empirical and simulated data. We see that the tails of empirical distributions for all three systems are well emulated by the ones obtained from the model. However, there are deviations which are the most likely consequence of using median values of parameters p_a, p_g, and p_{aff}.

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

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

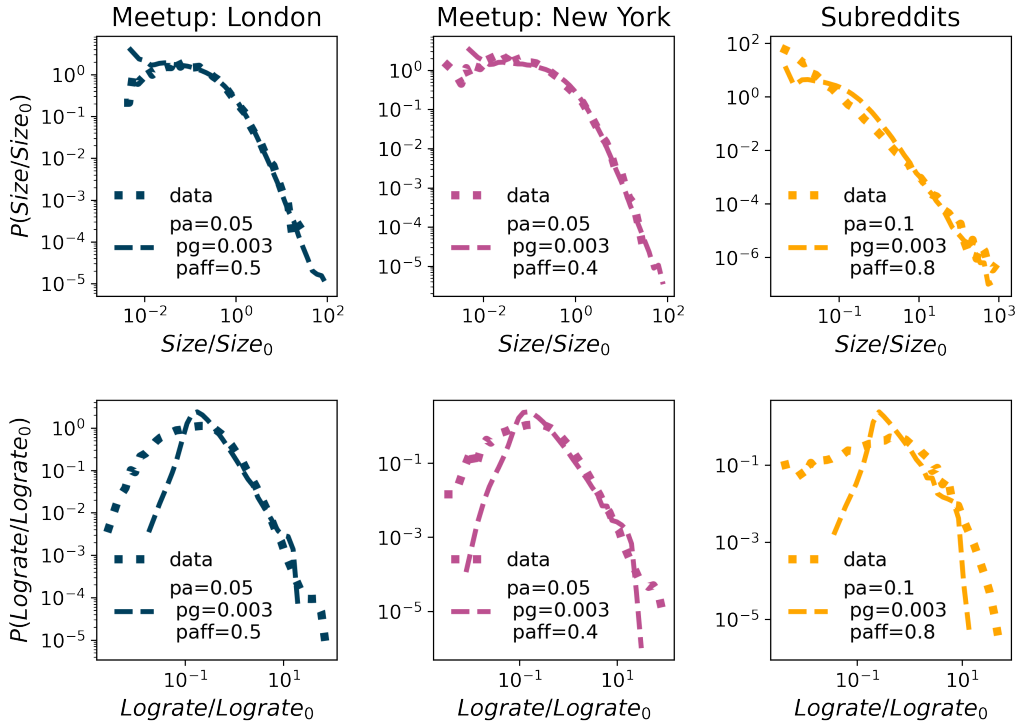

Figure 3.6: The comparison between empirical and simulation distribution for group sizes (top panel) and logrates (bottom panel).

3.6 Discussion and conclusions

The growth of the cities and companies has attracted the attention of researchers in the previous few decades [54, 52, **?**]. It is not surprising if we keep in mind that their growth determines other processes essential for the functioning of cities and economies. In the cities, economic and innovation growth scale with their size [**?**]. The understanding of growth and segmentation of economic systems are essential for the long-term prediction of their evolution, and risk assessment []. The growth of social groups and social system segmentation has been slightly overlooked since the focus was mainly on the structure of social networks and their evolution.

The results of our empirical analysis and theoretical modeling show that there are universal growth rules that govern the growth of social groups in these systems. Through rigorous empirical analysis of the growth of social groups in three systems, Meetup groups located in London and New York, and

Reddit, we show that the distribution of group sizes in these systems has log-normal normal behavior. The empirical distributions of normalized sizes of the groups created in different years fall on top of each other and have the same values of parameters for the same system. Furthermore, the distributions for Meetup groups located in London and New York have similar model parameter values, suggesting that groups' growth in these two systems are similar. Numerical simulations further confirm these findings. By tuning our model's parameters, we can reproduce the distribution of group sizes in all three systems.

Our results show that while the processes that govern the growth of social groups in studied social systems are the same, their importance varies among systems. The analysed groups grow through two mechanisms [51]: members join a group that is chosen according to their interests or social relations with the group's members. The number of members in the system is growing, as well as the number of groups. The empirical distribution of growth rates differs for Meetup and Reddit. The observed differences can be explained by different modalities of interactions between their members. Meetup members need to invest more time and resources to interact with their peers. The events are localised in time and space, and thus the influence of peers in selecting another social group may be limited. On the other hand, Reddit members do not have these limitations. The interactions are online, asynchronous, and thus not limited in time. The influence of peers in choosing new subreddits and topics thus becomes more important. The inspection of numerical simulations confirms these observations. The values of p_{aff} parameters for Meetup and Reddit imply that social connections in diffusion between groups are more critical in Reddit than in Meetup.

Gibrat's law is the first empirical law used by researchers to describe and explain the growth and segmentation of various socio-economical systems, including cities and firms. The possibility of application of common law to the growth of social groups in different systems indicates the existence of universal growth patterns and mechanisms that govern that growth []. Detailed and rigorous empirical analysis of the growth of the cities and firms showed that it goes beyond Gibrat's law []. Our and the work of other researchers [51] confirm that these findings also hold for the growth of social groups. The analysis of monthly growth rates shows that these rates are log-normally distributed and depend on the size of a group. Furthermore, we cannot reduce the model proposed in this work to the law of proportional growth. Although our analysis shows that Gibrat's law does not apply to the growth of social groups, our findings confirm that universal patterns characterise this growth.

The results presented in this paper contribute to our knowledge of the growth and segmentation of socio-economical systems. Our rigorous analysis shows that the distribution of sizes of groups for studied systems follows a log-normal distribution. The findings of the previous research suggested the power-law behavior of this distribution. A detailed and comprehensive analysis of distributions of group sizes in social systems is needed. These and future results will help us better understand the growth and segmentation of social systems and predict their evolution and sustainability.

3.7 Distributions fit

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

Table 3.2 summarizes the findings for empirical data on group size distributions from Meetup groups in London, Meetup groups in New York and Reddit. Using the maximum likelihood method, we obtain the parameters of the distributions [57]. The results indicate that log-normal distribution is the best fit for all three systems. Figure 3.7 shows the distributions of empirical data as well as log-normal fit on data. For Meetup data, we present fit on stretched exponential distribution, which very well fits a large portion of data. For subreddits, distribution is broad and, potentially, resembles power-law. Still, log-normal distribution is a more suitable fit.

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

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

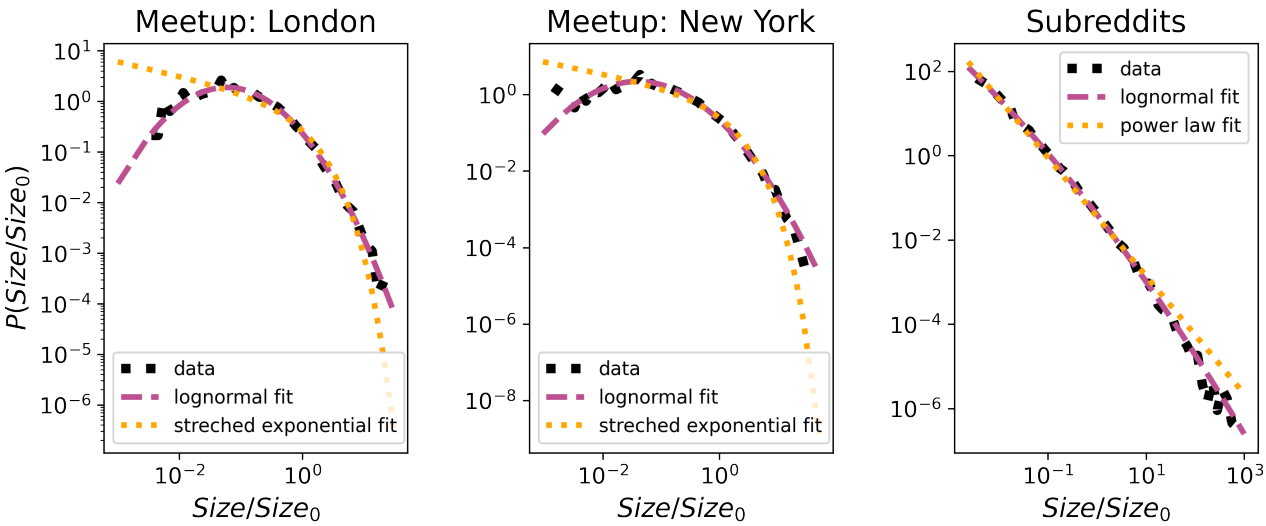

Figure 3.7: The comparison between log-normal and stretched exponential fit to London and NY data, and between log-normal and power law for Subreddits. The parameters for log-normal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits $\mu = -5.41$ and $\sigma = 3.07$.

We use the same methods to estimate the fit for simulated group size distributions on Meetup groups in London, New York, and Subreddits. Table 3.3 shows the results of the log-likelihood ratio R and p-value between different distributions. We conclude that log-normal distribution is most suitable for simulated group size distributions. Plotting log-normal and stretched exponential fit on data, Fig. 3.8 we confirm our observations.

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

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

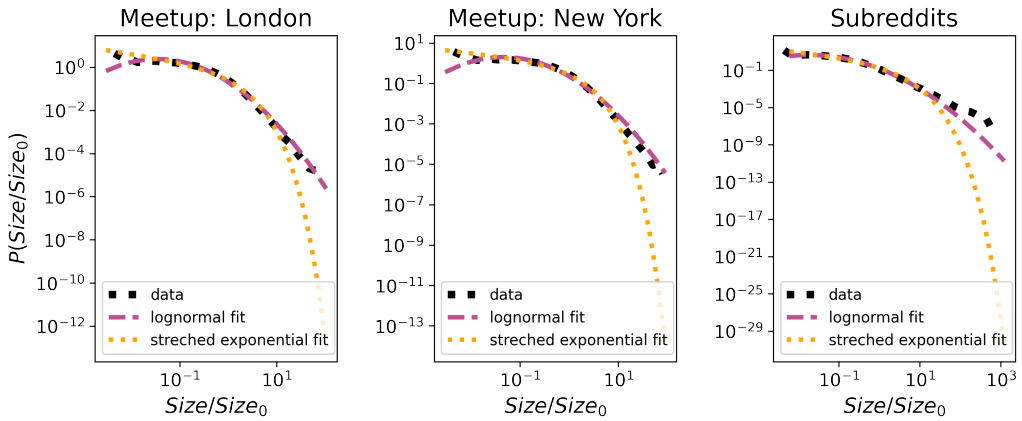

Figure 3.8: The comparison between lognormal and stretched exponential fit to simulated group sizes distributions. The parameters for log-normal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma = 1.53$.

3.8 The model for social groups growth

In the groups growth model, at each step, new users join the network, while old users are active with probability p_a. Active users can create new group with probability p_g. Otherwise, with probability p_{aff}, they perform diffusion linking. With probability, $1-p_{aff}$ users join a random group. Figure 3.9, top row, shows that group sizes distributions follow log-normal distribution. The affiliation parameter p_{aff} influences the width of distributions, so for larger p_{aff}, we find larger groups in the network. If, instead of random linking, users with probability $1 - p_{aff}$, choose to join to larger groups, group sizes distribution change significantly. Similar to affiliation model [51], group sizes have power-law distribution, see bottom row on Figure 3.9.

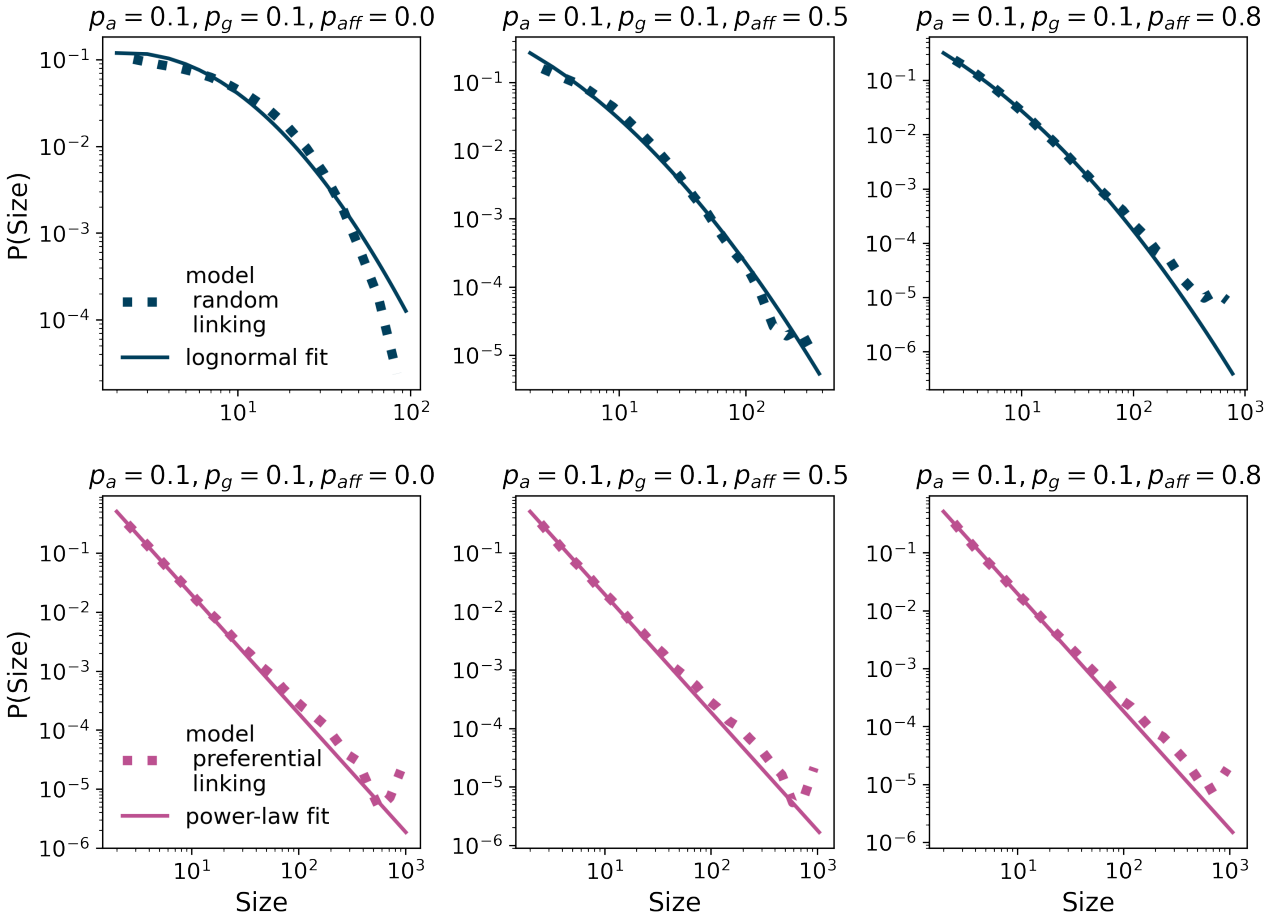

Figure 3.9: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff}. With probability, $1 - p_{aff}$, users choose a group randomly. The group sizes distribution (top row) is described with a log-normal distribution. With higher affiliation parameter, p_{aff}, distribution has larger width. The bottom row presents the case where with probability $1 - p_{aff}$ users have a preference toward larger groups. For all values of parameter p_{aff}, we find the power-law group sizes distribution.

Chapter 4

The role of trust in knowledge based communities

Information and communications technologies (ICTs) have enabled faster and easier creation and sharing of knowledge. Furthermore, they have provided access to a large amount of data which enabled a detailed study of their emergence and evolution [63], as well as user's roles [65], patterns of their activity [66, 67, 68]. However, relatively small attention was given to sustainability of SE communities. Most of the research was focused on the activity and factors that influence the increase of the users' activity in these communities. Factors such as need for experts and the quality of their contributions have been thoroughly investigated [69]. It was shown that growth of communities and mechanisms that drive it may depend on the topic around which the community was created [70].

4.1 The Stack Exchange

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

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

We focused analysis on four pairs of SE communities with the same topic. Astronomy, Literature

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

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

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.

4.1.1 Data

Stack Exchange data are public and regularly released. As closed communities were active between 180 and 210 days, we extracted only first 180 days of data. Given that first few months can be crucial for further development of the community [71], we are interested in early evolution of Stack Exchange sites.

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

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

Even though posts can be voted and downvoted, information about a user who voted is absent, so we do not consider these interactions between users. Comments can not be downvoted, while we find only around 3% negatively voted answers and questions, Table 4.1.

Table 4.1: 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%

4.1.2 Comparison between active and closed SE communities

Table 4.2 compares the first 180 days between closed and active communities. When it comes to basic statistics, active communities had larger number of users, questions, answers and comments. Another simple indicator if community is going to graduate or decline can be time series of active questions for period of 7 days in Figure 4.1. The question is active if had at least one activity, posted answer ot comment during previous seven days. We find that live communities have larger number of active questions after first three months. Still, this difference is smaller for literature and astronomy.

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

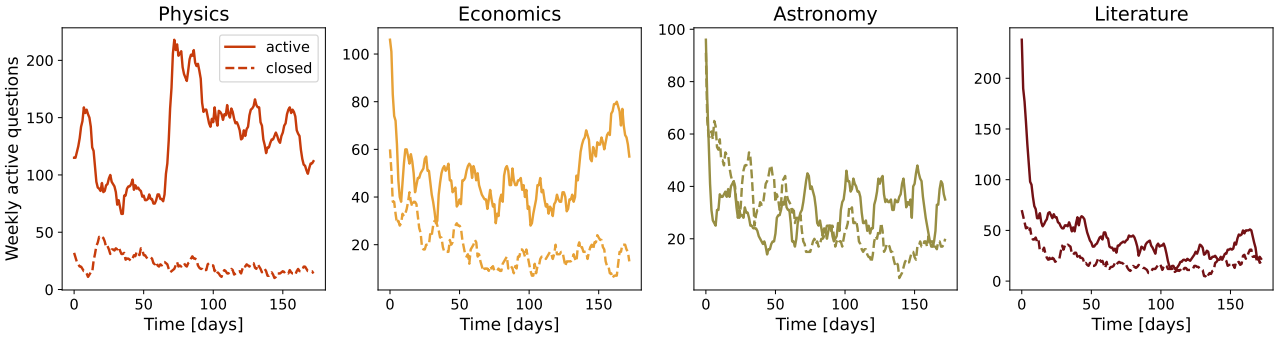

Figure 4.1: Number of active questions within 7 days sliding windows. Solid line - active sites; dashed lines - closed sites.

For astronomy we observe that closed community had higher number of active questions in the early period of community life.

Table 4.2: 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 [4]. To determine the the success of sites they measure how many questions are answered, how many questions are posted per day, and how many answers are posted per question. There are two measures: the number of avid users and the number of visits that are not easily interpreted from the data. The site is *healthy* if it has 10 questions per day, 2.5 answers per question and more than 90% of answered questions. For less than 80% of answered questions, 5 questions per day and 1 question per answer site *needs some work*.

We calculated Stack Exchange statistics for astronomy, economics, literature and physics and results are presented in the Table 4.3. After observed period of 180 days only live physics is healthy site while other live communities are at least in two criteria labeled as *okay*. Closed sites mostly *need some work*, the exception is closed astronomy. For example it has *excellent* percent of answered questions and *okay* answer ratio.

This simple measurements presented in tables 4.2 and 4.3 and in figure 4.1 do not provide us clear indications about community sustainability. Only for physics topic the difference between active and closed community is evident, while for other communities it is not so clear. Thus, we need deeper

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

Table 4.3: 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

insights into structure and dynamics of these communities to understand. The structure of social interactions within communities and dynamics of collective trust may provide better explanation why some communities succeed and other died.

4.2 Network properties of Stack Exchange data

On Stack Exchange sites, the interaction between users happens through posts. As we are interested in examining the characteristics of the users, we map interaction data to the networks. Using complex network theory, we can quantify the properties of obtained networks and compare different SE communities, e.g. alive and closed SE sites. In the user interaction network, the link between two nodes, user i and j, exists if user i answers or comments on the question posted by user j, or user i comments on the answer posted by user j. The created network is undirected and unweighted, meaning that we do not consider multiply interactions between users or the direction of the interaction.

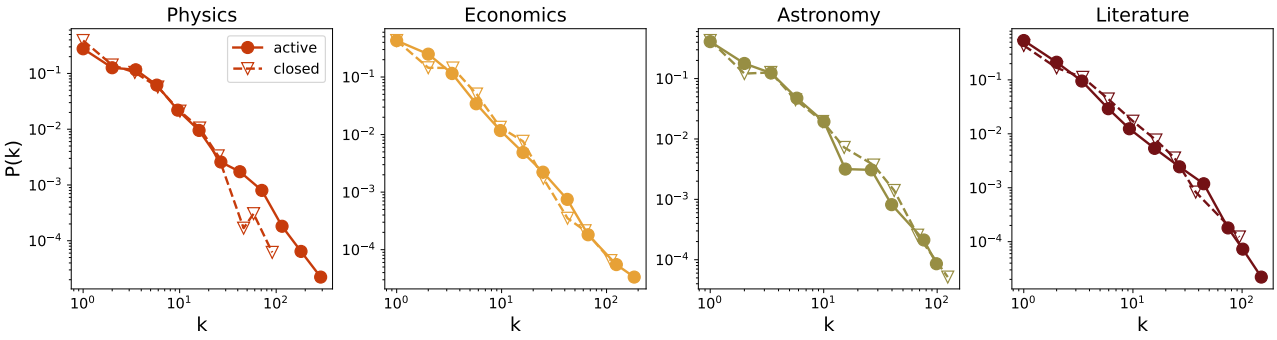

Figure 4.2: Degree distribution.

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

We calculate different structural properties of observed networks and study their evolution. There are many local and global measures of the network properties [1]. They are also dependent, still it

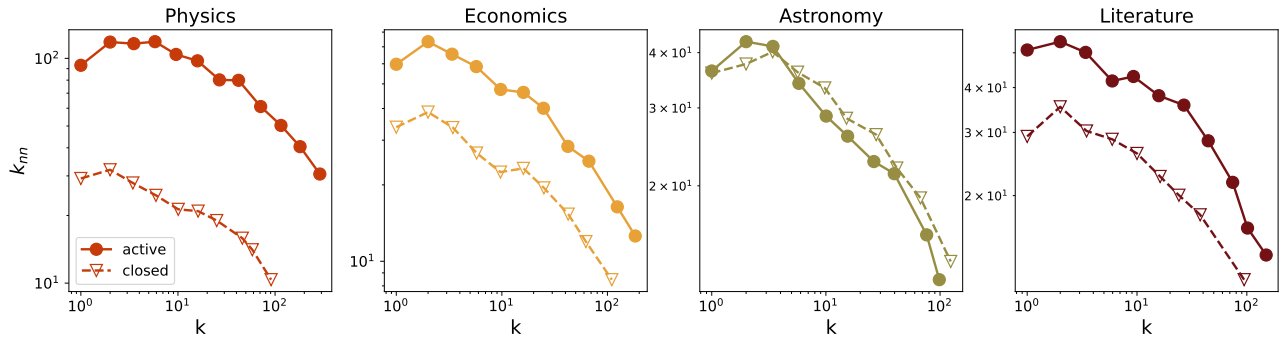

Figure 4.3: Neighbour degree.

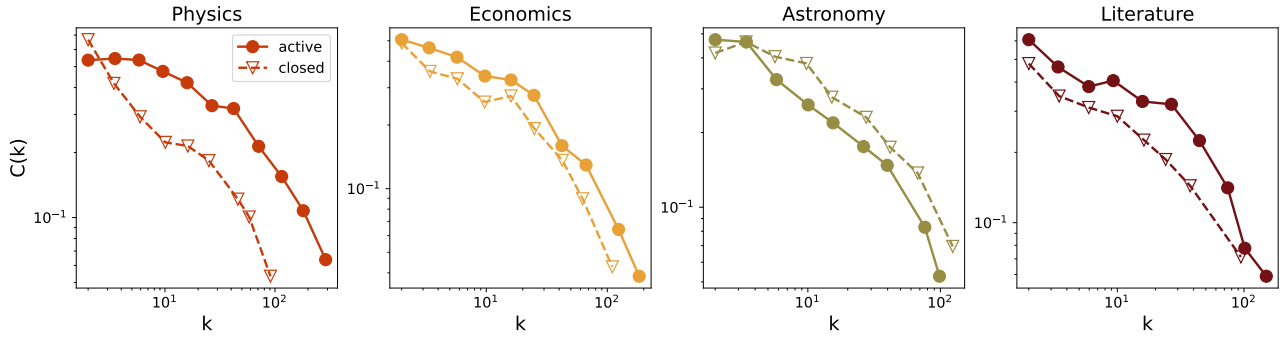

Figure 4.4: Neighbour degree.

was shown that degree distribution, degree-degree correlations and clustering coefficient can describe the properties of the most complex networks [72].

4.2.1 Clustering

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

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

Here e_i is the number of links between neighbours of the node i in a network, while $\frac{1}{2}k_i(k_i - 1)$ is the maximal possible number of links determined by the node degree k_i. The clustering coefficient of network C is the value of clustering averaged over all nodes. Here we investigate how clustering coefficient in a SE community is changing with time by calculating its value for all network snapshots. We compare the behavior of clustering for active and closed communities on the same topic in order to better understand how cohesion of these communities is changing over time. Members' clustering coefficient measures the probability that other members connected to them are also connected. Study on dynamics of social group growth shows that that links between one's friends that are members of a social group increase the probability that that individual will join the social group [48]. Furthermore, successful social diffusion typically occur in networks with high value of clustering coefficient [73]. These results suggest that high local cohesion should be a characteristic of sustainable communities.

We first analyse structural properties of Stack Exchange communities and examine the difference between successful and unsuccessful ones. We calculate the mean clustering coefficient for 30-days window networks and examine how it changes with time. Figure 4.5 shows the evolution of mean

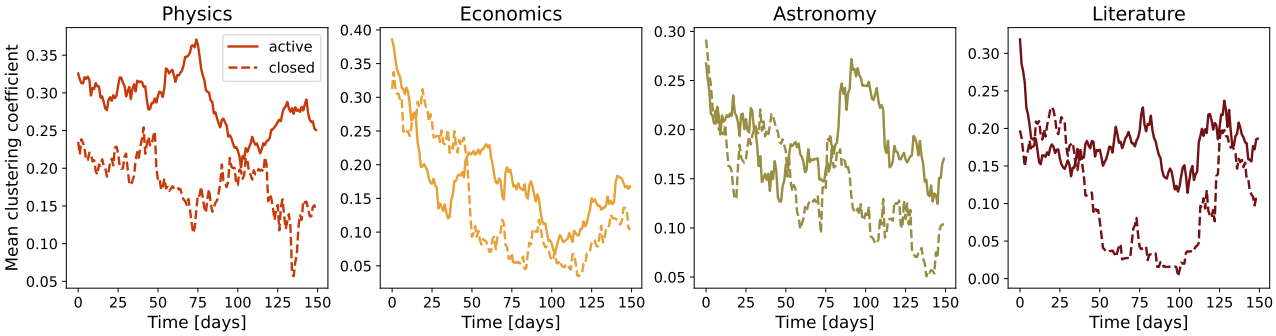

Figure 4.5: Mean clustering coefficient.

clustering coefficient for all eight communities. All communities that are still alive are clustered, with the value of mean clustering coefficient higher than 0.1. Physics, the only launched community, has the value of clustering coefficient above 0.2 for the first 180 days. During larger part of the observed period, the clustering coefficient of an active community is higher compared to the clustering coefficient of its closed pair. If we compare active communities with their closed counterpart, the closed communities have higher value of the mean clustering coefficient in the early phase while later communities that are still active have higher values of clustering coefficient. These results suggest that all communities have relatively high local cohesiveness, and that lower values of clustering coefficient in the later phase of community life may be an indicator of its decline.

4.2.2 Core-periphery structure

Real networks, including social networks, have a distinct mesocopic structure [74, 19]. Mescoscopic structure is manifested either through community structure or core-periphery structure. Networks that have community structure consist of a certain number of group of nodes that are densely connected with each other, with sparse connections between groups. Networks with core-periphery structure consist of two groups of nodes, with higher edge density within one group and between groups but low edge density in the second group [19]. Research on dynamics of user interaction in SE communities shows that there is a small group of highly active members, similar to core, that have frequent interactions with casual or low active members of community [66, 70]. These results indicate that we should expect a core-periphery structure in SE communities.

Core-periphery pattern means that network consists of two components: a core, densely connected group of nodes, and periphery, a second group of nodes that are loosely connected with the core and with each other. Classification of nodes into one of these two groups provide us with information about their functional and dynamical roles in the network. Active users typically belong to core, while periphery consists of less active users.

To investigate core-periphery structure of SE communities and how it evolves trough time, we analyse the core-periphery structure of every 30 days network snapshot. We use Stochastic Block Model (SBM) adapted for core-periphery inference of network structure [19] to determine the core-periphery structure.

For each 30 days snapshot network we run 50 iterations and choose the model parameters θ and p according to minimum description length. MDL does not change much among inferred core-periphery structure, see Fig. A8, while looking into adjusted rand index we can notice that difference exists. Still, ARI between pair-wised compared partitions is large ($ari > 0.9$) indicating stability of inferred structures.

4.2.3 core-periphery structure of knowledge-sharing networks

Furthermore, we examine core-periphery structure of these communities and its evolution. Specifically, we are interested in the evolution of connectivity in the core. Figure 4.6 shows the number of links between nodes in the core per node $\frac{L}{N}(t)$. $\frac{2L}{N}$ is the average degree of the node in the core, and thus, $\frac{L}{N}$ is the half of the average degree. Again, Physics community has the much higher value of this quantity than Theoretical physics during the whole observed period, indicating higher connectivity between core members. Higher connectivity between core members in the active community is also characteristic for Literature, although this quantity has the same value for active and closed communities at the end of the observation period. The differences between active and closed communities are not that evident for Economics and Astronomy, see Fig. 4.6. Active and closed Economics communities have similar connectivity in the core during the first 50 days. After this period, the connectivity in the core of the active community the twice as large as in the closed community and the difference grows at the end of observation period. The connectivity in the core of closed Astronomy community is higher that the connectivity in the core of the active community during the first 50 days. But as the time progresses, this difference changes in the favor of live community, while at the end of the observation period the difference disappears.

The difference between active and closed communities is more prominent if we consider average number of core-periphery edges per core node. The connectivity between core and periphery is higher for the still active communities than for the closed ones, see Fig. 4.6. This is very obvious if we compare Physics and Theoretical physics community. Moreover, Physics community has the highest connectivity compared to all other communities. When it comes to active communities that are still in the beta phase, they either have the same core-periphery connectivity as their closed counter part, or as in the case of Astronomy, their periphery is weaker connected to the core during the first 50 days of their life, see Fig. 4.6.

On average, the cores of the active communities have higher number of nodes in the core than the closed communities, Fig. A11. However, the relative size of the core compared to the size of the whole network is similar when we compare closed and active communities on the same topic. This is even true for communities on physics topic. The size of the core fluctuates with time for active and closed communities. The core membership also changes with time. This core membership is changing more for the closed communities. We quantify this by calculating the Jaccard index between the cores of the subnetworks in the moment t_i and t_j. Figure A9 in Supplementary Information shows the value of Jaccard index between any two of the 150 subnetworks. The highest value of the Jaccard index is around the diagonal and has value close to 1. This is expected, since these subnetworks are for consecutive days and the difference between them is smaller. The value of Jaccard index decreases with number of days between two subnetworks $|t_i - t_j|$ faster in closed communities Fig. A10. This difference is the most prominent for the literature communities, while this difference is practically non-existent for Astronomy. The relatively high overlap between cores of even more distant subnetworks for active communities, further confirms that the core is more stable in these communities that in their closed counterparts.

4.2.4 Core-periphery structure of the interaction networks

In Q-A communities are common two types of users: popular and casual users. Popular users tend to generate the majority of interactions - they are likely to post more questions, also take part in answering questions and tend to engage discussions through comments. For popular users we consider 10 of most active users. We analyse interactions between popular and casual users and among popular users in the sub-networks of 30 days [t+30). In both cases the number of links per nodes in active sites are larger than in closed communities (figure 4.11).

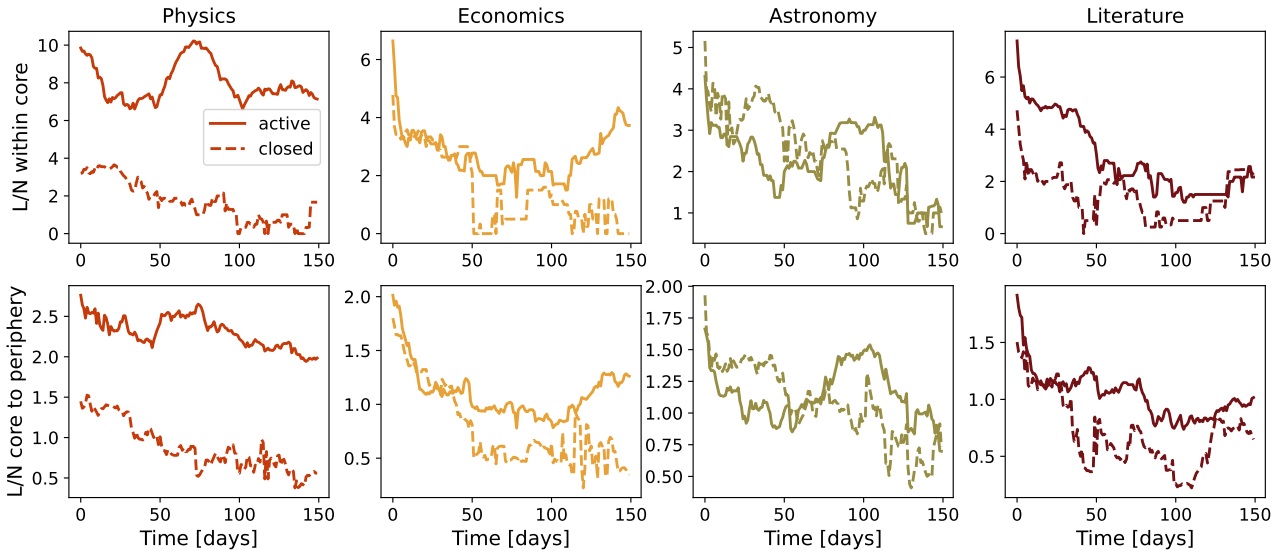

Figure 4.6: Links per node in core and links per node between core and periphery.

Although this separation of users puts an emphasis on differences between closed and active sites, it does not guarantee that all popular users are in the top 10 . To solve this dilemma we use the SBM (Stochastic Block Model) algorithm) to detect the core and the periphery of each 30 days sub-network. Such a split of users leads us to similar conclusions as before. (see figure A.3 - 2nd column)

Stochastic models start from random configuration and the algorithm can converge to different local stable states. For each 30 days sub-network we run 50 iterations of SBM and choose the model parameters θ, p according to minimum description length. As example we show analysis of inferred sample of core-periphery structures for 30 days area51 astronomy networks, Figure B.1. We represent mean minimum description length (MDL) and mean number of nodes in the core with standard deviation. MDL does not change much among inferred core-periphery structures, still difference between obtained configurations is notable in the number of nodes in the core. To investigate in more details similarity between obtained core-periphery configurations in the sample we calculate several measures between pair-wised partitions such as normalized mutual information, adjusted rand index, F1 measure and Jaccard index. Those measures are larger than 0.5, and in most cases higher than 0.9 indicating stability of the inferred core-periphery structures.

To study the stability of the core across the time we compute Jaccard's coefficient between core users in [t+30) networks selected at times t_1 and t_2, (figure 4.7). Higher values of the Jaccard index indicate that core users tend to stay in the core. The detected cores experience a lot of change over time and the highest overlap between core users is in the network closer in the time. The average Jaccard index between core users in all sub-networks separated by time interval $|t_1 - t_2|$ with standard deviation confidence interval is presented in figure 4.8. Compared to closed sites, active sites show more stability in the core. Even the number of core users obtained in the launched and closed communities are comparable 4.9 (a high difference is found only for physics), the ratio between total core and periphery reputation is evidently higher in the active than in closed sites, figure 4.10.

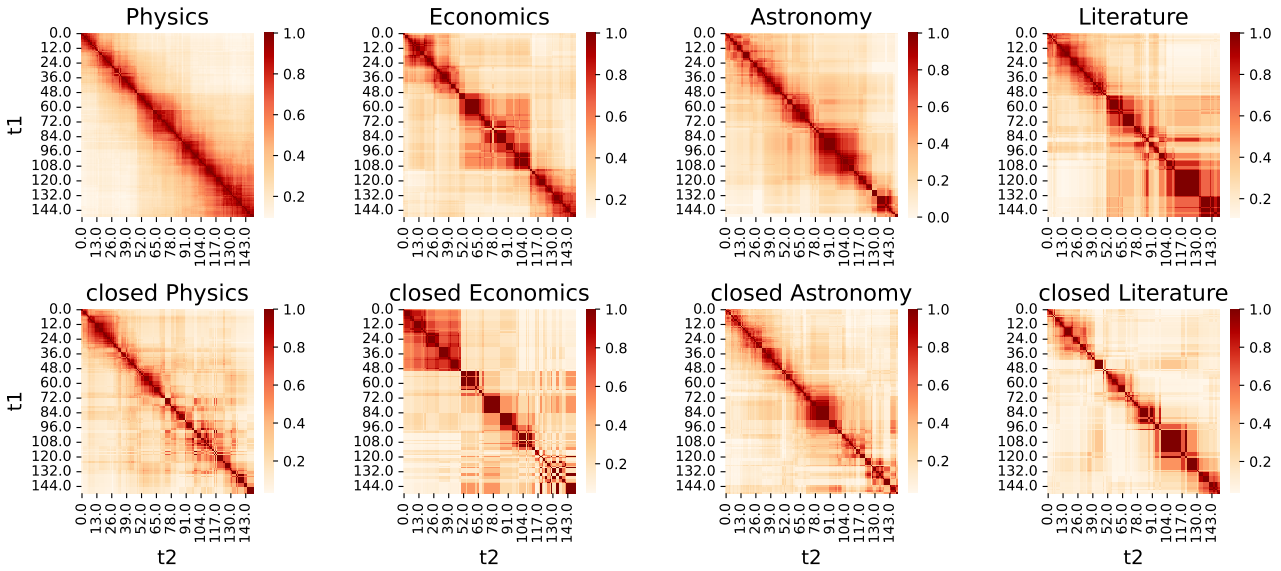

Figure 4.7: Jaccard index between core users in sub-networks at time points $t1$ and $t2$

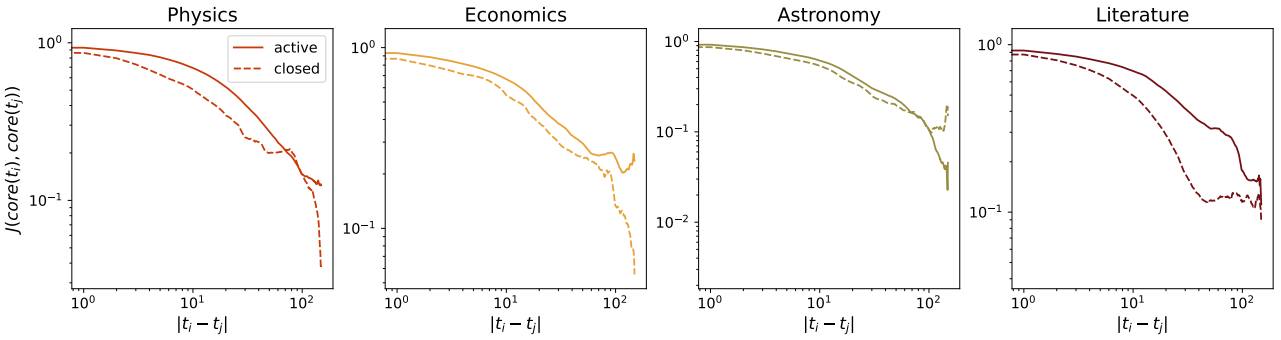

Figure 4.8: Jaccard index between core users in 30days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$

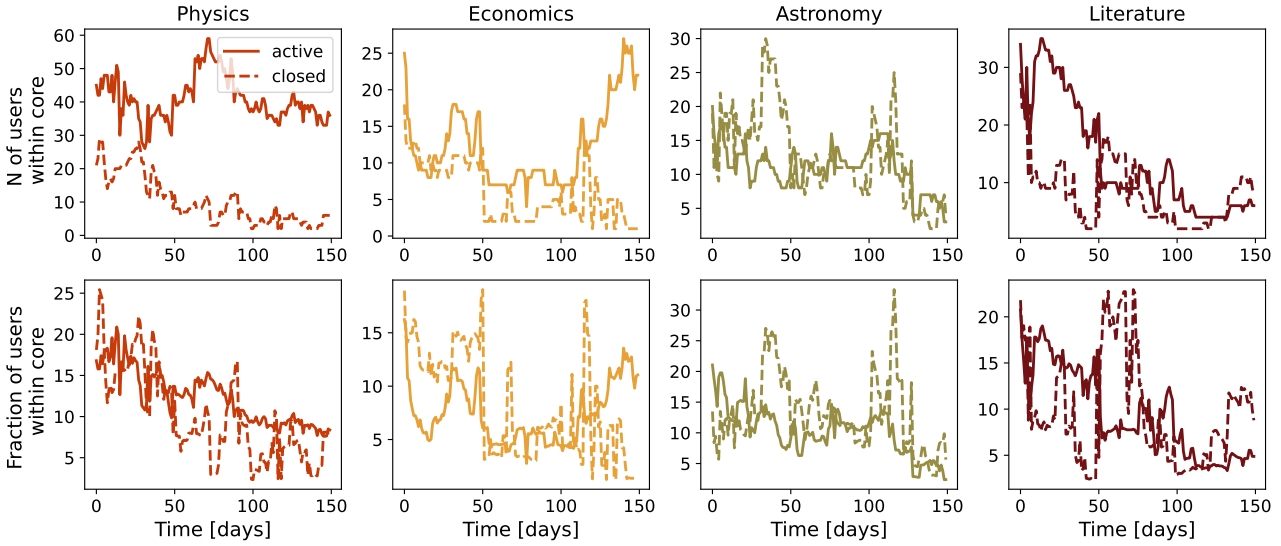

Figure 4.9: Just for reference size of the core (top) and fraction of users in core (bottom). Solid lines - active sites; dashed lines - closed sites.

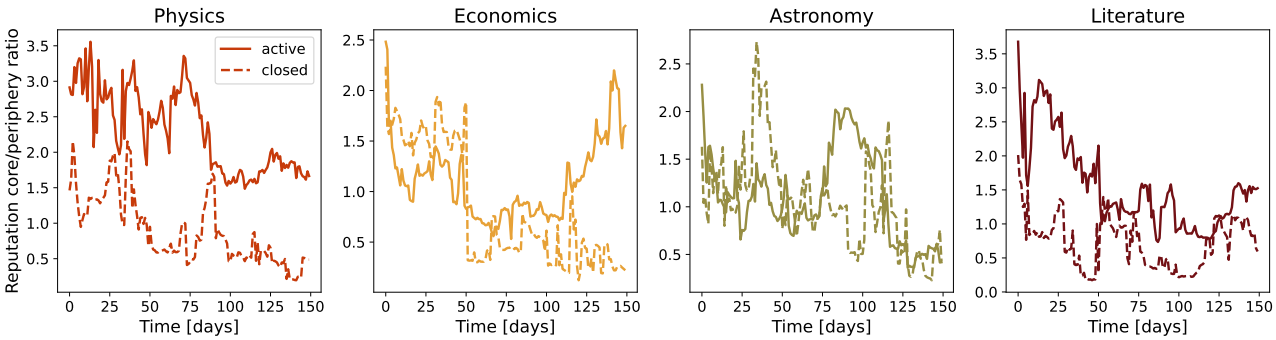

Figure 4.10: Ratio between the total reputation within network core and periphery. Solid lines beta communities, dashed lines area 51 communities.

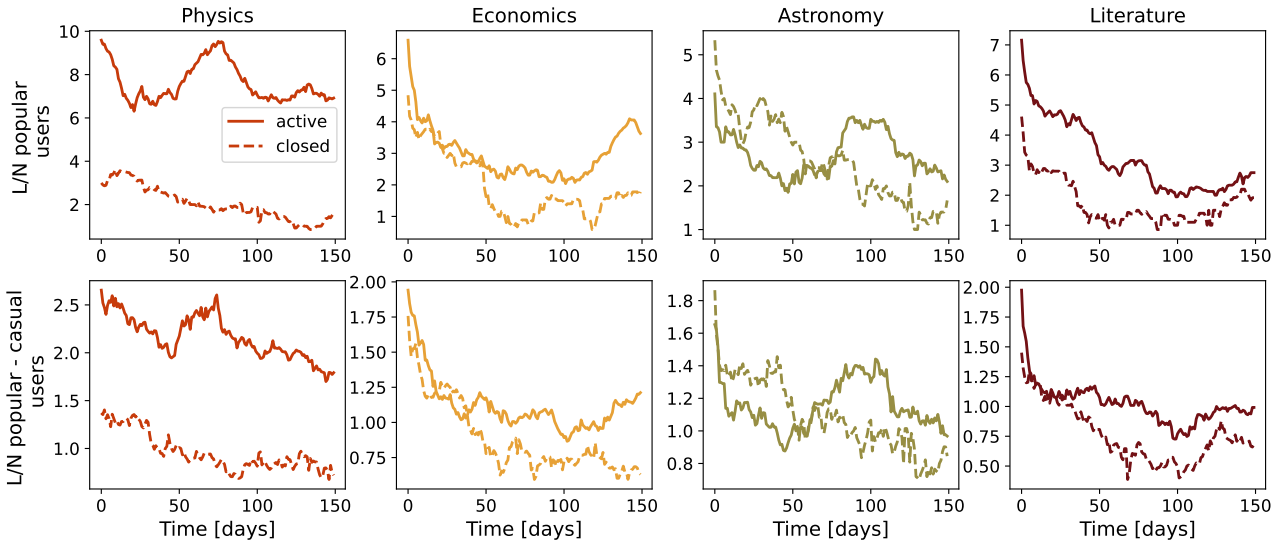

Figure 4.11: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users). Reminder: only 3rd and 5th columns should stay and only for reference to previous research, while our point is to this selection via core/periphery decomposition without thresholding.

4.2.5 Dynamical Reputation model

4.2.6 Selection of dynamical reputation model parameters

One of the largest drawbacks of DIBRM is the parameter tuning problem. In previous applications of the model [27, 75] there was no single best set of parameter values for modeling dynamic reputation in Stack Exchange communities. For example, in [75] the best approximation of the official Stack Exchange reputation is obtained with $t_a = 2, \beta = 1, \alpha = 1.4$ which means there is no active forgetting factor. In our application of DIBRM to SE communities we opted for a different set of parameter values. Details of parameter search and tuning are presented in SI.

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

Figure A2 in SI shows the difference between the number of users that had at least one activity in the window of 30 days and number of users with reputation higher than 1 during the same period for different values of parameter β. The minimal difference between these two variables is observed for the values of β between 0.94 and 0.96 for both live and closed communities. Since we want to compare communities, we select $\beta = 0.96$ after verifying that this level of reputational decay does not reduce the number of active users (based on their dynamic reputation) below the actual number of users who have been active (interacted with the community) in the time window of 30 days.

To summarize, our model of dynamical reputation has three parameters: 1) basic reputation contribution $I_{bn} = 1$; 2) cumulative factor $\alpha = 2$; 3) forgetting factor $\beta = 0.96$. The selected values of parameters are used for measuring dynamical reputation of user in all four pair SE communities. Given these values of parameters, the minimal reputation achieved by the user immediately after they have made an interaction in the SE community is 1. This reputation will decay below 1 if the user does not perform another interaction within the one-day time window. For any user in a community, when their reputation drops below 1, we consider this user inactive which means that the user at that time is not "visible" in the community and their past contributions at that time are unlikely to impact other users. The number of active users and mean user reputation for different Stack Exchange communities are shown in Fig. 4.17.

4.2.7 Dynamic reputation - β parameter

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

In Dynamic reputation model, the β parameter controls the strength of the forgetting fator of the model. The value of this parameter should reflect the core feature of the reputational systems and make reputation easier to loose. Due to user's inactivity, any level of reputation will eventually decay to below 1. Dependence of time needed for reputation to drop below this level and the β parameter, as well as reputation before inactivity is shown on Figure 4.12. Here I_n is equal to the raw number of interactions in the community without forgetting or cumulative factor at work.

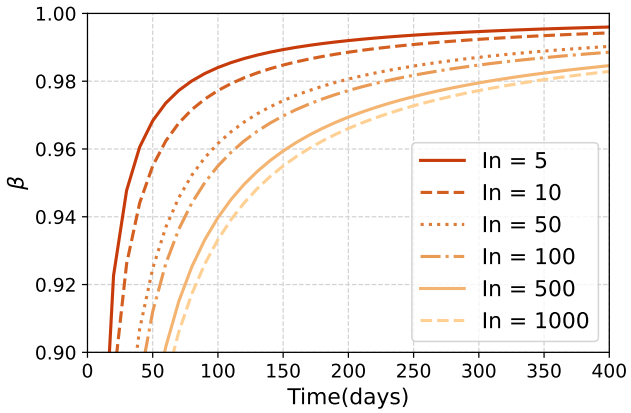

Figure 4.12: Dependence of parameter β and number of days Δ needed for reputation I_n to drop to $I_{n_0} = 1$. Dependence of parameter β and number of days when reputation due inactivity decreases from I_n to I_0 is given as $\beta = (\frac{I_{n0}}{I_n})^{(1/\Delta)}$

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

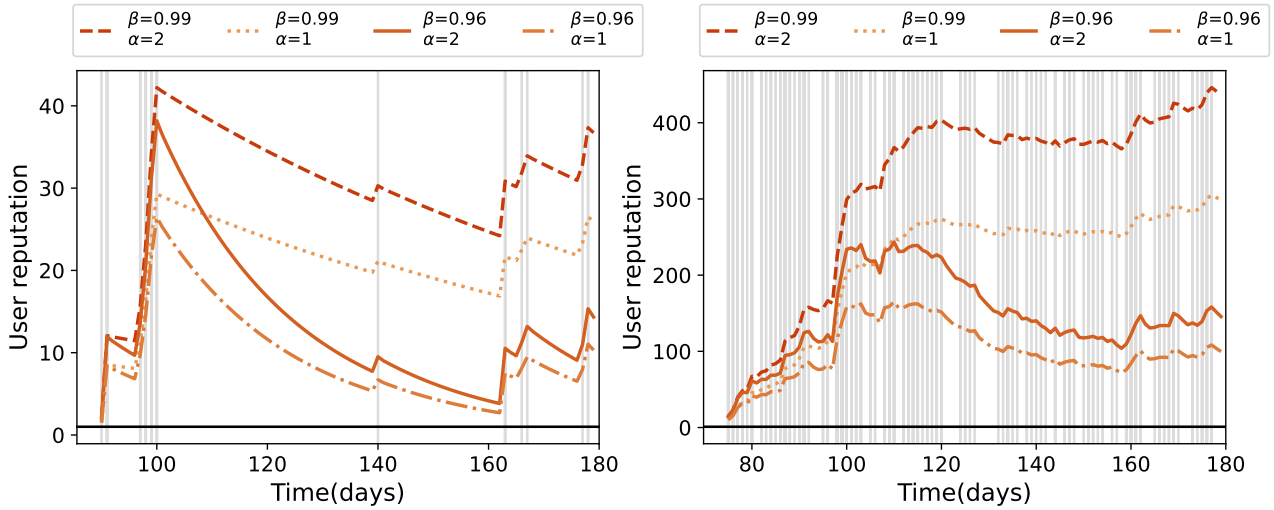

Figure 4.13: Single users reputations

and around five months for high reputation based on 500 or 1000 interactions to decay below the threshold.

30 days sliding window We compared the number of users with estimated reputation higher than 1 for different parameters β and concluded that β close to 0.96 approximates the number of users with recorded interactions in a given 30 days sliding window. For each pair of communities we calculated number of users with at least one interactions in every 30 days sliding window and then we estimated several times series expressing the number of users with reputation higher than 1 for fixed β. Then we calculated the root mean square error (RMSE) between those time series for the first 200 days. Values of RMSE are shown on Figure 4.14. 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.

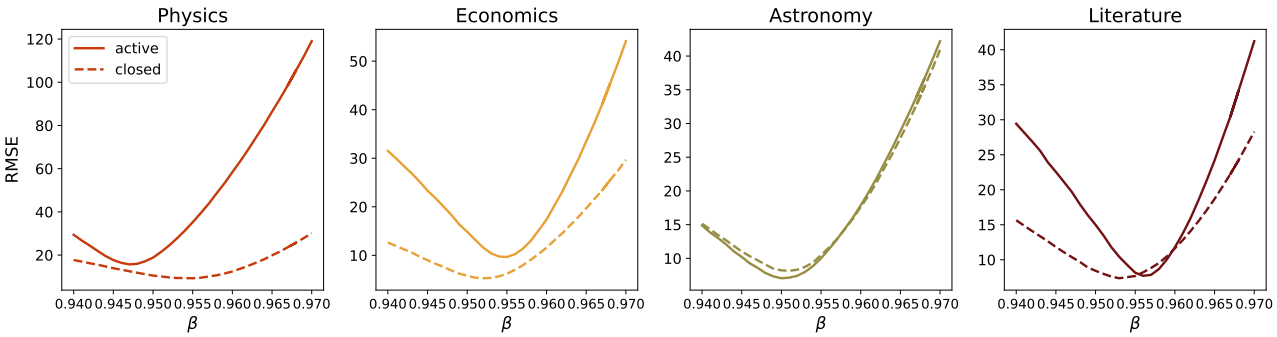

Figure 4.14: RMSE between number of active users in sliding window of 30 days and number of users with reputation > 1 for $0.94 < \beta < 0.97$ with step 0.001.

Figure 4.15 shows comparison between number of users in 30 days sliding window, number of users for these optimal values $\beta = 0.954$ and $\beta = 0.96$. For $\beta = 0.96$ we observe that in most communities estimated number of active users consistently slightly higher than the actual number of users which have made at least one interaction in that sliding window. This means that dynamic reputation model in some cases overestimates the reputation of the user, but far more important is that it never understimates the real number of active users. Since we base our calculations of total and average reputation within the community only on users whose reputation is higher than the threshold this is important as no active users are disregarded by the model due to the value of the decay parameter.

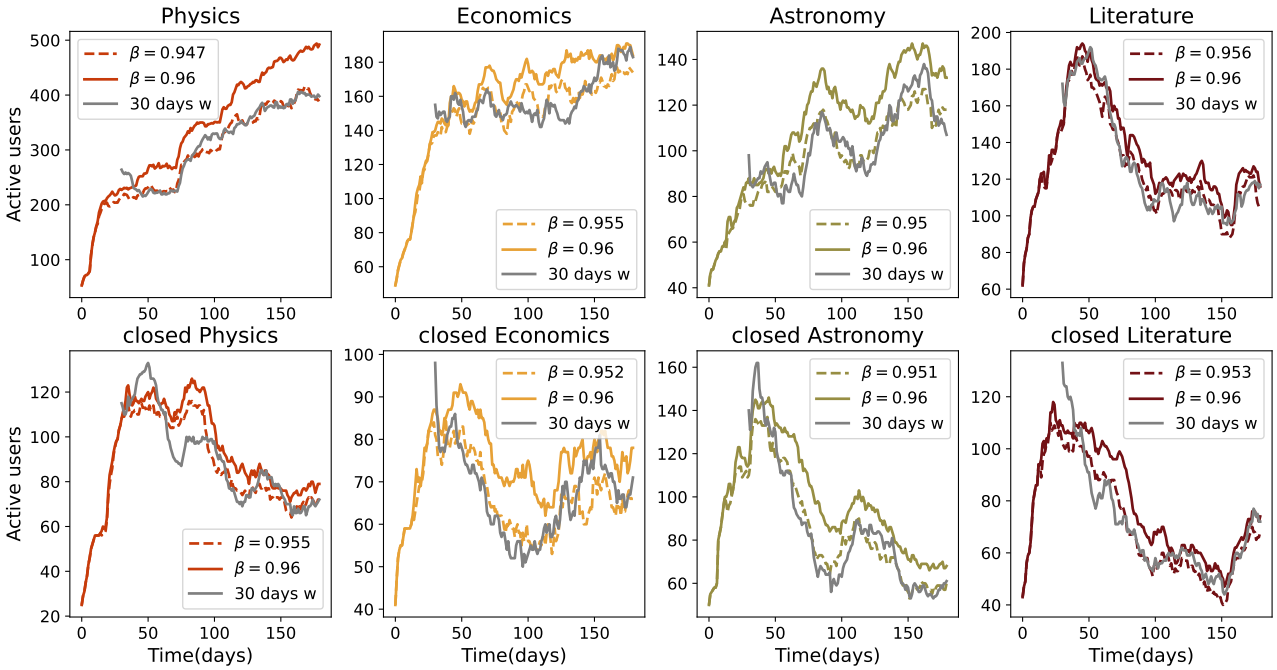

Figure 4.15: Number of active users in a sliding window of 30 days and number of users with dynamic reputation higher than 1 for $\beta = 0.954$ and $\beta = 0.96$ which provide the best fit to the number of users in 30 days sub-networks for each community

Finally, it's imporant that our dynamic reptuation captures the trend of long-term user activity. In Figure 4.16 solid lines show the time series of estimated dynamic reputation for $\beta = 0.96$ while dashed lines show the number of users who were active in a given sliding window and continued to be active in the next one. Although the total estimated number of active users is expectedly higher, two time series follow similar trends in different communities.

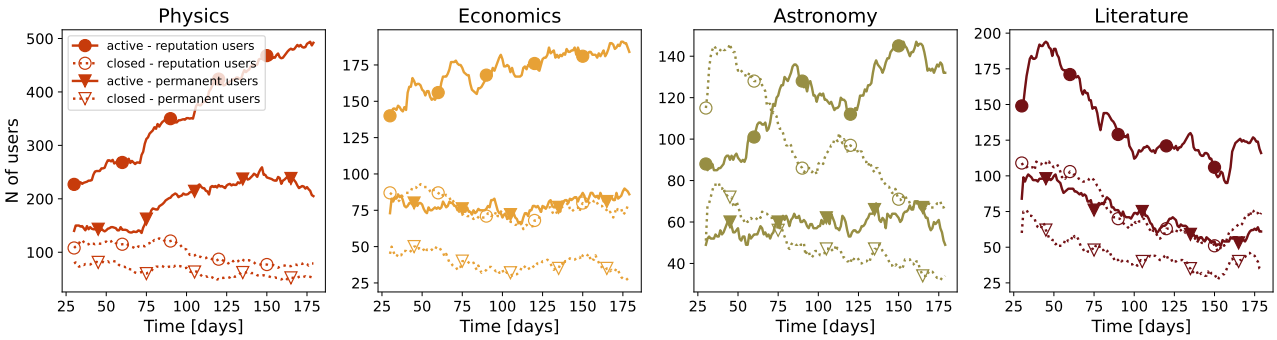

Figure 4.16: Solid lines represent number of users with dynamic reputation higher than 1 for $\beta = 0.96$ while dashed lines are number of users within 30 days sliding window who were active and remained to be active. Blue lines are beta, while red lines are area51 communities.

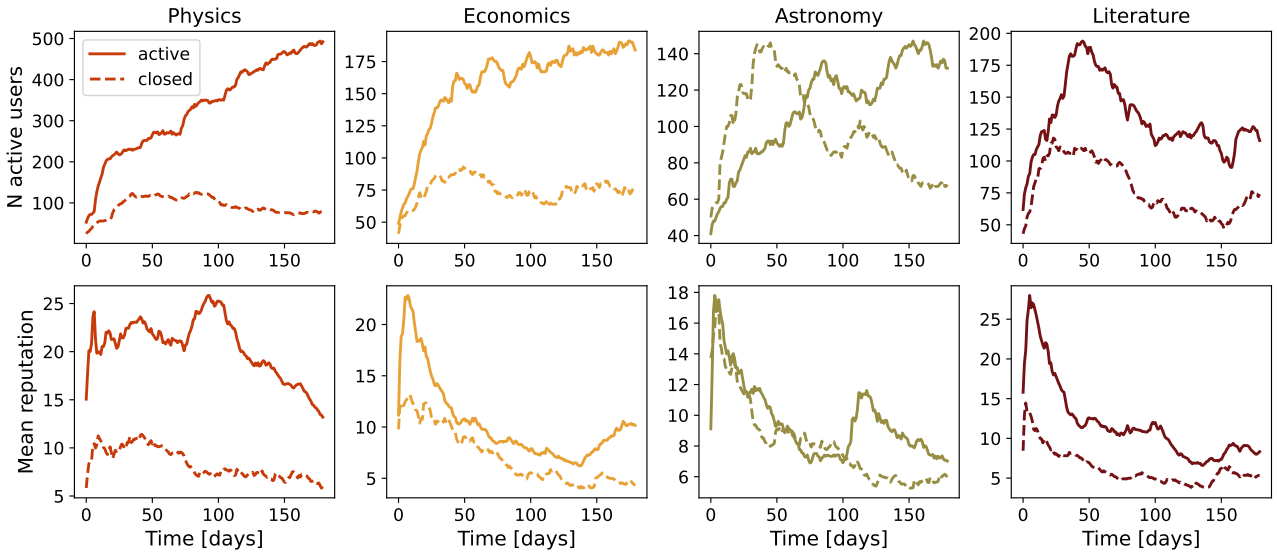

Figure 4.17: Dynamic Reputation on the four pairs of Stack Exchange websites: Astronomy, Literature, Economics, Physics and Theoretical Physics.

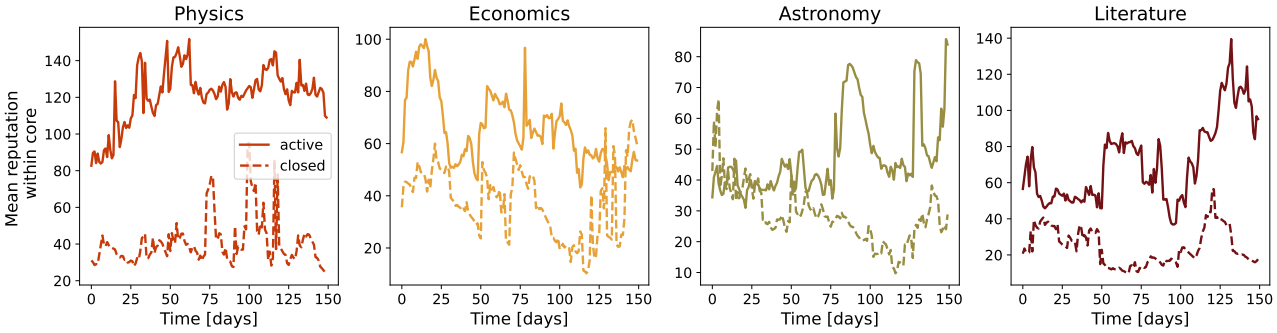

Figure 4.18: Dynamical reputation within core.

4.2.8 Dynamic reputation of users within the network of interactions

Examined network properties suggest that there are structural differences between active and closed communities. Active communities have higher and more stable local cohesiveness compared to their closed counterparts. The overlap of the set of nodes in the core for active communities shows a significant overlap even for distant subnetworks, meaning that the membership of the core in active communities is more stable.

To further explore the differences between active and closed communities, we focus on dynamical reputation which is our proxy for collective trust in these communities. We investigate whether and how core-periphery structure is related to collective trust in the network. Figure 4.18 shows the mean dynamical reputation in the core of active and closed communities and its evolution during the observation period. There are clear differences between active and closed communities when it comes to dynamical reputation. The mean dynamical reputation of core users is always higher in active communities than in closed. As expected, the largest difference is observed between Physics and Theoretical Physics community. The difference between active communities which are still in the beta phase and their closed counterparts is not as prominent, however, the active communities have higher mean dynamical reputation especially in the later phase of community life. The only difference in the pattern is observed for astronomy communities at the early phase of their life, when closed community has a higher value of dynamical reputation than active community. This is in line

with similar patterns in the evolution of mean clustering and core-periphery structure.

By definition, the core consists of very active individuals and thus we expect higher total dynamical reputation of users in the core in comparison to the the total reputation of users belonging to subnetworks periphery. Figure A12 shows the ratio between the total reputation of core and periphery for closed and active communities and its evolution. The ratio between total reputation of core and periphery in Physics is always higher than in the Theoretical physics community. Similar pattern can be observed for literature communities, although the difference is not as clear as in the case of physics. Ratio of total dynamical reputation between core and periphery is higher for closed community than active one on the economics topic in the early days of community life. However, in the later stage of their lives this ratio becomes higher for active communities. Communities around astronomy topic deviate from this pattern, which once again shows the specificity of these communities.

To complete the description of the evolution of dynamic reputation active and closed communities, we examine the evolution of Gini index of dynamical reputation in the whole network which is shown in Fig. A5 in Supplementary Information. The Gini index is always higher for active communities than for closed ones, especially for later times in observation period. Only pattern of Astronomy communities deviates from the pattern observed for other three pairs during the early days. These results indicate that the dynamical reputation is distributed in the population more unequally in the active than in closed communities. The evolution of assortativity coefficient that measures correlations between dynamical reputation of connected users in the subnetworks, shown in Fig. A6, shows that networks are disassortative for the largest part of the observation period. These results suggest that users with high dynamical reputation have tendency to connect with users with low value of dynamical reputation.

In Figure **??** we show mean user reputation in core and in periphery over time (30 day sliding windows as before). We see that the mean user reputation in core is greater in the currently active sites (solid lines, top panels) than in their closed pairs (dashed lines). In the bottom panels, we see that the mean reputation on the network periphery has substantially lower values, and the difference between active and closed sites is less pronounced.

For reference in Fig 4.9 we show core sizes in all sites. We show these in absolute numbers (total number of nodes) and as a fraction of network size through time.

Gini coefficient Besides the number of active users (who at given moment of observation have reputation higher than the threshold) and the population mean value of dynamical reptuation, we have investigated in more details the distribution of dynamical reputation within discussed communities. We have observed that the distributions are often skewed which prompted us to compare the communities in terms of their Gini coefficient. The gini coefficient is a simple measure that shows us the degree of reputation inequality within the community. We calculate the value based on the dynamic reputation values of users at every time step (day) and report he values in Fig. 4.19. We see that all communities (both still active and closed ones) have gini coeffiecinet values higher than 0.5 throughout first six month period. Interestingly, except in the case of Astronomy, currently active communities had higher reputation inequality every day during first six month period. As in many other measures, in the case of astronomy, closed community started as more unequal one (signalled by higher gini coef values), but after around two months the situation changed.

4.2.9 Dynamic reputation in the network of interactions

In the few figures below, we investigate whether users' dynamic reputation is related with users' position within the network.

Dynamic Reputation assortativity

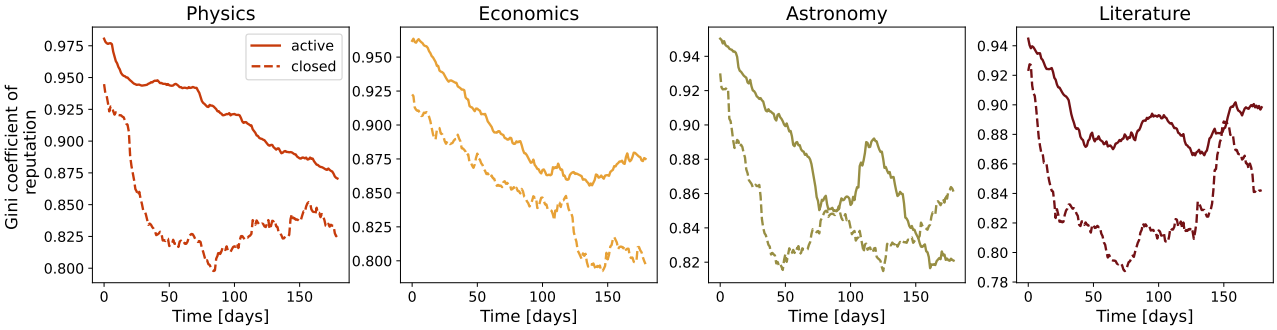

Figure 4.19: Gini index of dynamic reputation within population

We first look at user interaction patters, e.g. we investigate whether users connect with others of similar or different reputation (positive/negative assortativity). We operationalize this by measuring assortativity of dynamic reputation on interaction network. Practically this is a meassure of correlation between dynamic reputation of users who are linked in the interaction network. These results are shown in Fig. 4.20. We look at 30 day unweighted undirected networks of interactions (questions, answers and comments) and calculate assortativity by using users' reputation on the last day of observed time window. We see small values of assortativity that are mostly negative, signaling weak correlations between reputation levels of interacting users. The fact that the values are mostly negative are expected, users of different dynamic reputation interact, e.g. active, high reputation users respond to the questions of new, less reputable users. Exceptions are closed astronomy and literature sites that occasionally had positive assortativity values, signaling existence of links between users of similar reputation levels.

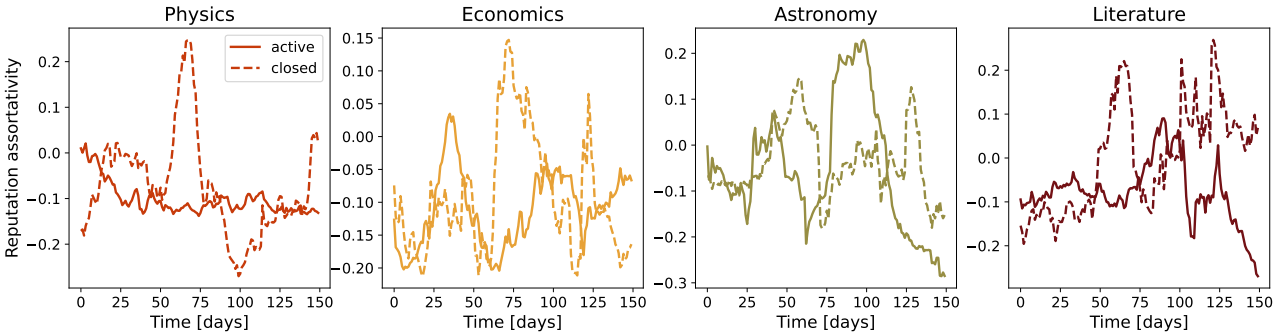

Figure 4.20: Dynamic Reputation assortativity in the network of interactions (questions, answers, comments, unweighted, undirected network). Solid lines - active sites; dashed lines - closed sites.

DynRep & Degree DynRep & BC

We continue to investigate whether the user's reputation correlates with typical network centrality measures calculated at user's node in the interaction network. As previously, we compare node's centrality in the 30 day network with the node's dynamic reputation on the last day of the period, repeat the process every day for the first six months. Correlation coefficient between dynamic reputation and degree in the network is very high, as expected, as most of the interactions that contributed to user's reputation are also present as links in the network. We show these results in Fig. 4.21(top). However, we again see the distinction between active and closed communities where this correlation is higher in active communities, except in the first month of sliding windows. Astronomy is an exception here as well as we see that the correlations were similar in both closed and still active sites throughout observed period. In the bottom panels of Fig. 4.21 we present correlation coefficients of dynamic

reputation and user's betweenness centrality in the interaction network. These corrlations are also high and most of the time higher in the later networks of active than closed communities. This is particularly interesting due to global nature of betweenness centrality measure and less obvious relation of it to user's dynamic reputation.

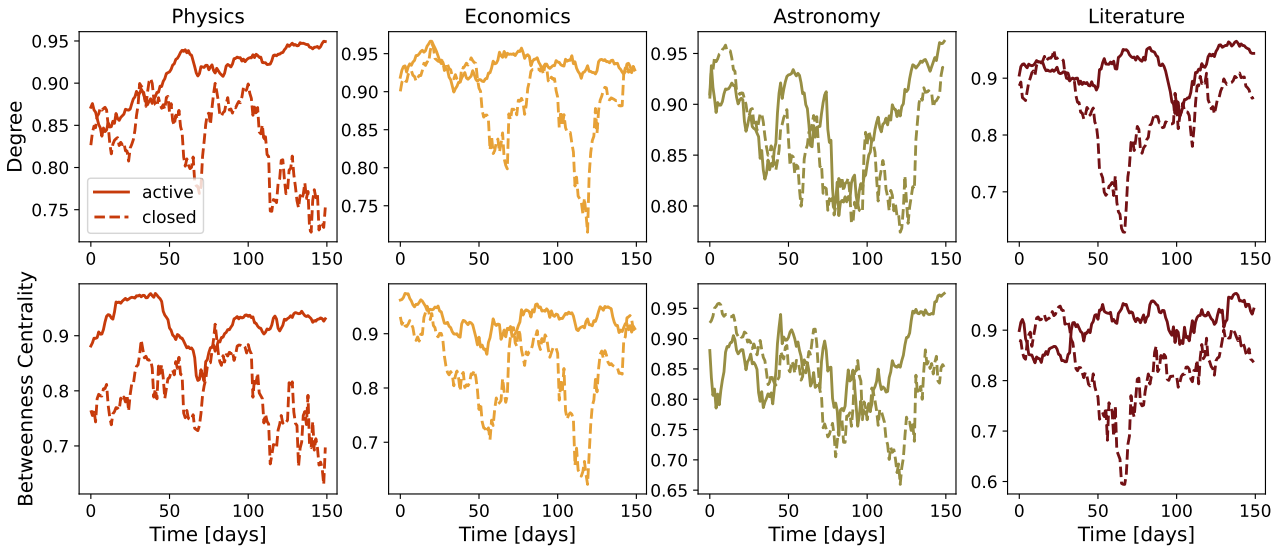

Figure 4.21: Coefficient of correlation between users' Dynamic Reputation and users' network degree (top) and users's betweenness centrality (bottom). Solid lines - active sites; dashed lines - closed sites.

Appendix A

The choice of the sliding window

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

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

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

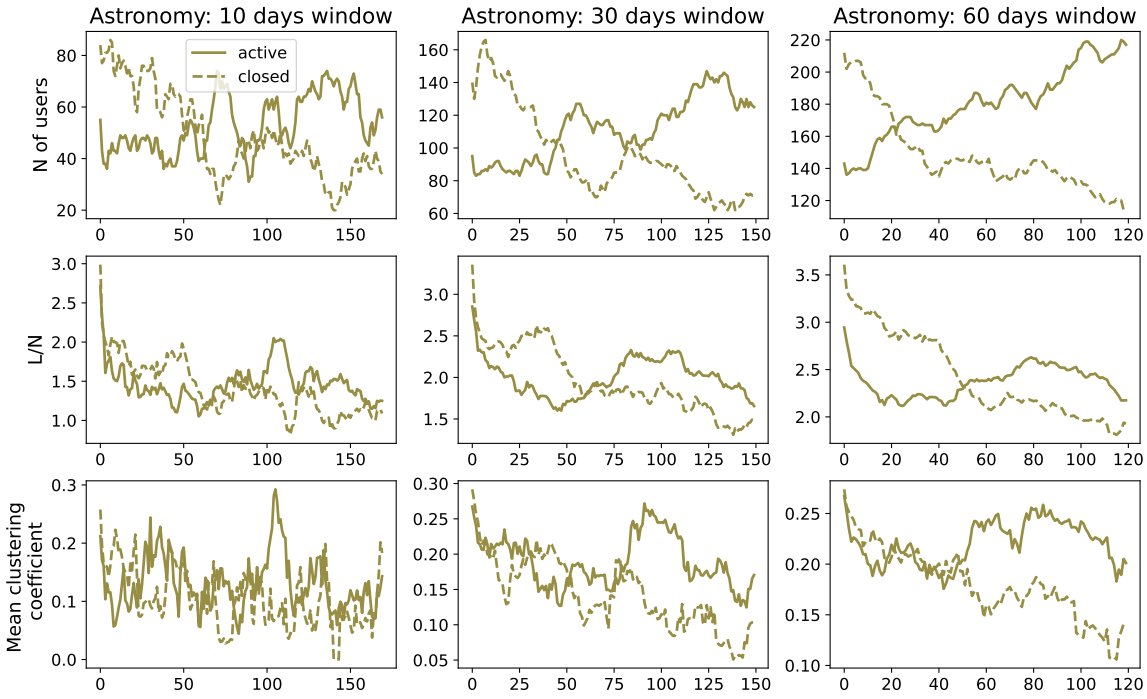

Figure A.1: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

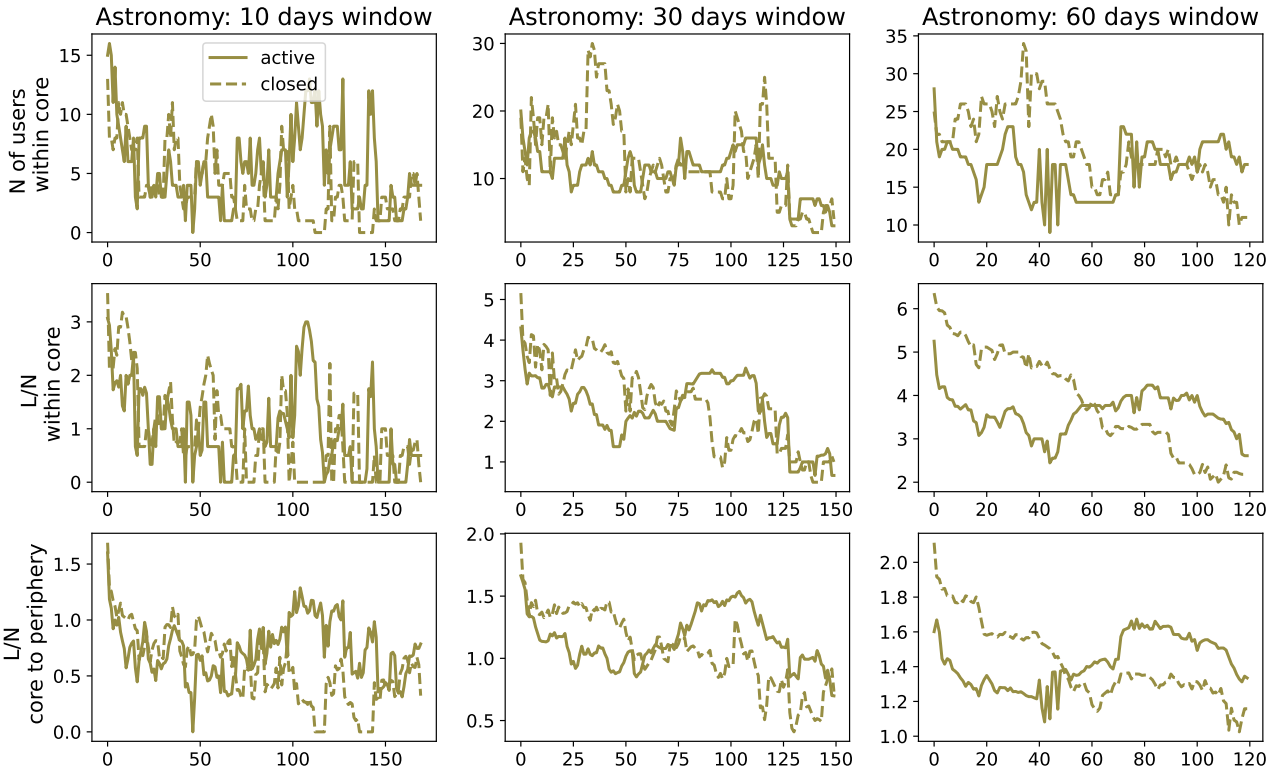

Figure A.2: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

between large and small time scales.

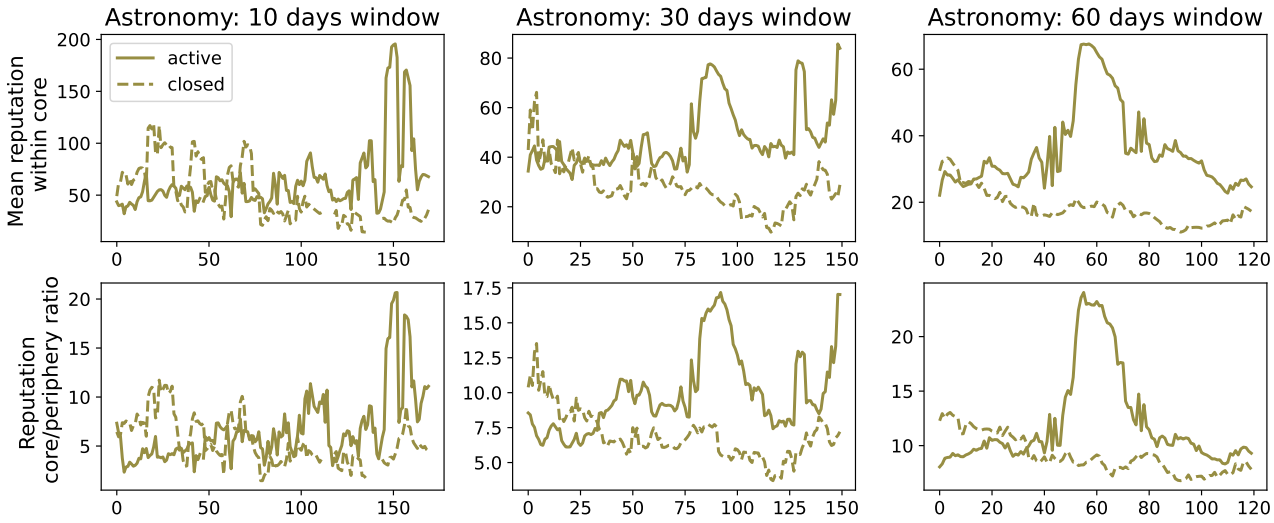

Figure A.3: Results for different sliding windows. Example is for astronomy, blue solid lines- active, orange dashed lines - closed site.

Appendix B

Robustness of core-periphery algorithm

B.0.1 Precision and recall

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

The core-periphery structure has two groups so confusion matrix [76] 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) [76].

B.0.2 F1 measure

The **F1 measure** is the harmonic mean of precision and recall [76]:

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

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

B.0.3 Jaccard coefficient

The **Jaccard's** coefficient is the ratio of two classes' intersection to their union [76]. It can also be expressed in terms of confusion matrix:

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

B.0.4 Normalized mutual information

Normalized mutual information (NMI) is similarity measure between to partitions C and $C^{'}$ based on information theory [77]:

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

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:

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

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.

B.0.5 Adjusted rand index

Adjusted rand index. For the set of nodes V, with n nodes, consider all possible combination of pairs (v_i, v_j). We can select the number of the pairs where nodes belong to the same group in both partitions, C and $C^{'}$, denoted as a. Similarly, as b, we can define the number of pairs whose nodes belong to different groups in partitions. Then, unadjusted rand index [78] 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 [79] , defined as:

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

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

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

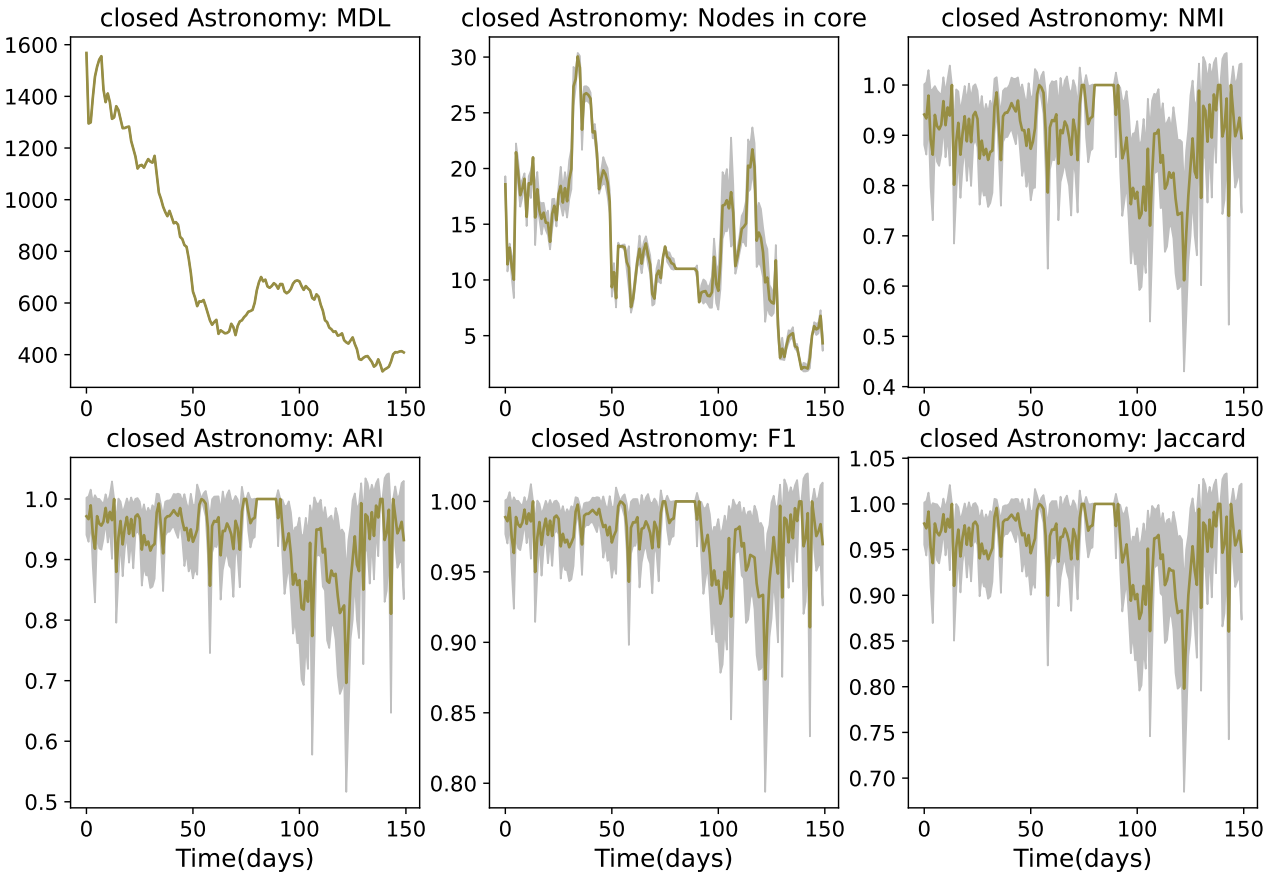

Figure B.1: Minimum description length, number of nodes in core, normalized mutual information, adjusted rand index, F1 measure and Jaccard index, among 50 samples for 30-days sub-networks. Results are given for closed astronomy.

Bibliography

[1] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.

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

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

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

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

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

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

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

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

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

[11] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):1–16, 2012.

[12] Naomi A Arnold, Benjamin Steer, Imane Hafnaoui, Hugo A Parada G, Raul J Mondragon, Félix Cuadrado, and Richard G Clegg. Moving with the times: Investigating the alt-right network gab with temporal interaction graphs. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–17, 2021.

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

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

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

[16] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M Pardalos, Cristina Masoller, and Martín G Ravetti. Quantification of network structural dissimilarities. *Nature communications*, 8(1):1–10, 2017.

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

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

[19] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *arXiv preprint arXiv:2005.10191*, 2020.

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

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

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

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

[24] Jan W Kantelhardt, Stephan A Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):87–114, 2002.

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

[26] Claudiu Duma, Nahid Shahmehri, and Germano Caronni. Dynamic trust metrics for peer-to-peer systems. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, pages 776–781. IEEE, 2005.

[27] 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.

[28] 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.

[29] 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.

[30] C-K Peng, Sergey V Buldyrev, Shlomo Havlin, Michael Simons, H Eugene Stanley, and Ary L Goldberger. Mosaic organization of dna nucleotides. *Physical review e*, 49(2):1685, 1994.

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

[32] 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.

[33] 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.

[34] Milovan Suvakov, Marija Mitrovic, Vladimir Gligorijevic, and Bosiljka Tadic. How the online social networks are used: dialogues-based structure of myspace. *Journal of The Royal Society Interface*, 10(79):20120819, 2013.

[35] Marc Barthelemy. *The structure and dynamics of cities*. Cambridge University Press, 2016.

[36] César A Hidalgo and Ricardo Hausmann. The building blocks of economic complexity. *Proceedings of the national academy of sciences*, 106(26):10570–10575, 2009.

[37] Jelena Smiljanić, Arnab Chatterjee, Tomi Kauppinen, and Marija Mitrović Dankulov. A theoretical model for the associative nature of conference participation. *PloS one*, 11(2):e0148528, 2016.

[38] Ali Montazeri, Soghra Jarvandi, Shahpar Haghighat, Mariam Vahdani, Akram Sajadian, Mandana Ebrahimi, and Mehregan Haji-Mahmoodi. Anxiety and depression in breast cancer patients before and after participation in a cancer support group. *Patient education and counseling*, 45(3):195–198, 2001.

[39] Kathryn P Davison, James W Pennebaker, and Sally S Dickerson. Who talks? the social psychology of illness support groups. *American Psychologist*, 55(2):205, 2000.

[40] Wendy K Tam Cho, James G Gimpel, Daron R Shaw, et al. The tea party movement and the geography of collective action. *Quarterly Journal of Political Science*, 7(2):105–133, 2012.

[41] Sinan Aral and Dylan Walker. Identifying influential and susceptible members of social networks. *Science*, 337(6092):337–341, 2012.

[42] Sandra González-Bailón, Javier Borge-Holthoefer, and Yamir Moreno. Broadcasters and hidden influentials in online protest diffusion. *American behavioral scientist*, 57(7):943–965, 2013.

[43] János Török, Gerardo Iñiguez, Taha Yasseri, Maxi San Miguel, Kimmo Kaski, and János Kertész. Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment. *Physical review letters*, 110(8):088701, 2013.

[44] Taha Yasseri, Robert Sumi, András Rung, András Kornai, and János Kertész. Dynamics of conflicts in wikipedia. *PloS one*, 7(6):e38869, 2012.

[45] Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591, 2009.

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

[47] 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.

[48] 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.

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

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Konglin Zhu, Wenzhong Li, Xiaoming Fu, and Jan Nagler. How do online social networks grow? *Plos one*, 9(6):e100023, 2014.

[56] 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.

[57] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLOS ONE*, 9(1):1–11, 01 2014.

[58] 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.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] Akrati Saxena and Harita Reddy. Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *Journal of Computational Social Science*, pages 1–33, 2021.

[66] 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.

[67] Rogier Slag, Mike de Waard, and Alberto Bacchelli. One-day flies on stackoverflow-why the vast majority of stackoverflow users only posts once. In *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, pages 458–461. IEEE, 2015.

[68] Anamika Chhabra and S RS Iyengar. Activity-selection behavior of users in stackexchange websites. In *Companion Proceedings of the Web Conference 2020*, pages 105–106, 2020.

[69] Himel Dev, Chase Geigle, Qingtao Hu, Jiahui Zheng, and Hari Sundaram. The size conundrum: Why online knowledge markets can fail at scale. In *Proceedings of the 2018 World Wide Web Conference*, pages 65–75, 2018.

[70] 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.

[71] 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.

[72] Chiara Orsini, Marija M Dankulov, Pol Colomer-de Simón, Almerima Jamakovic, Priya Mahadevan, Amin Vahdat, Kevin E Bassler, Zoltán Toroczkai, Marián Boguná, Guido Caldarelli, et al. Quantifying randomness in real networks. *Nature communications*, 6(1):1–10, 2015.

[73] 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.

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

[75] 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.

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

[77] 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.

[78] 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.

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