



Sustainability of Stack Exchange Q&A communities: the role of trust

Ana Vranić^{1*} , Aleksandar Tomašević², Aleksandra Alorić^{1,3} and Marija Mitrović Dankulov¹

*Correspondence: anav@ipb.ac.rs

¹Institute of Physics Belgrade,
University of Belgrade, Pregrevice
118, Belgrade, Serbia

Full list of author information is
available at the end of the article

Abstract

Knowledge-sharing communities are fundamental elements of a knowledge-based society. Understanding how different factors influence their sustainability is of crucial importance. We explore the role of the social network structure and social trust in their sustainability. We analyze the early evolution of social networks in four pairs of active and closed Stack Exchange communities on topics of physics, astronomy, economics, and literature and use a dynamical reputation model to quantify the evolution of social trust in them. In addition, we study the evolution of two active communities on mathematics topics and two closed communities about startups and compare them with our main results. Active communities have higher local cohesiveness and develop stable, better-connected, trustworthy cores. The early emergence of a stable and trustworthy core may be crucial for sustainable knowledge-sharing communities.

Keywords: Networks structure; Dynamic reputation; Knowledge exchange; Stack Exchange; Sustainability of Q&A communities

1 Introduction

The development of a knowledge-based society is one of the critical processes in the modern world [1, 2]. In a knowledge-based society, knowledge is generated, shared, and made available to all members. It is a vital resource. Sharing this resource between individuals and organizations is a necessary process, and knowledge-sharing communities are one of the fundamental elements of a knowledge society.

Often, these knowledge-sharing communities depend on the willingness of their members to engage in an exchange of information and knowledge. Participation in the community is voluntary, with no noticeable material gains for members. Recent research has shown that the process of knowledge and information exchange is strongly influenced by *trust* [3, 4]. The exchange of knowledge depends on trust between a member and the community. It is a collective phenomenon that depends on and is built through social interactions between community members. This is why we believe it is crucial to understand how trustworthy knowledge-sharing communities emerge and disappear, as well as to unveil the fundamental mechanisms that underlie their evolution and determine their sustainability.

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

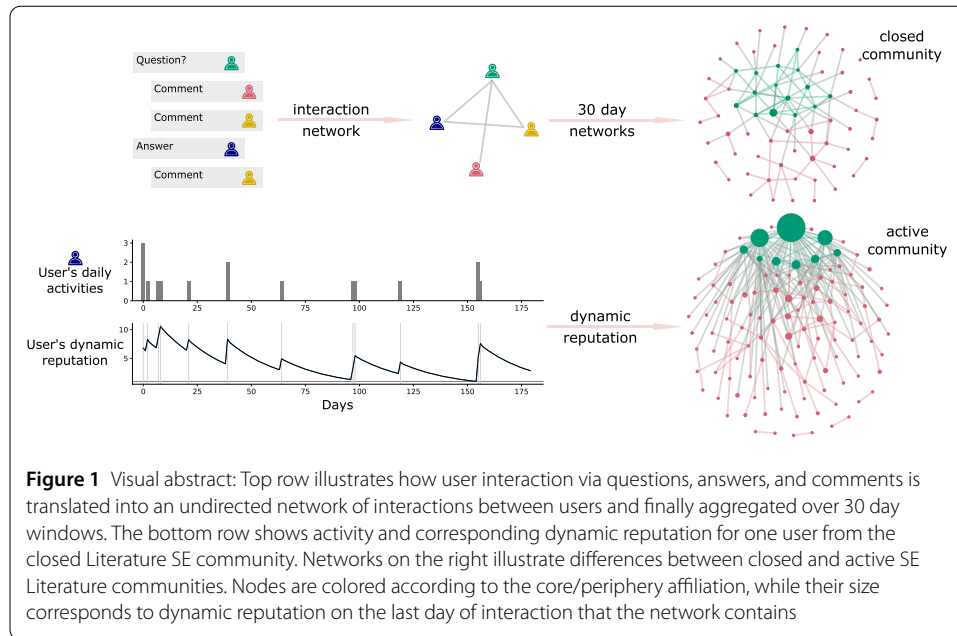
Unlike small offline knowledge-sharing groups, online communities consist of a large number of members where repeatable mutual interactions between all members are not possible. Thus, the trustworthiness of individuals in these communities has to be assessed and signaled using other means. It was shown that the reputation of an individual within the community is a strong signal of her trustworthiness that can override the main sources of social bias [5]. The reputation helps users manage the complexity of the collaborative environment by signaling out trustworthy members.

In the past two decades, we have witnessed the emergence of an online knowledge-sharing community Stack Overflow, which has become one of the most popular sites in the world and the primary knowledge resource for coding. The success of Stack Overflow led to the emergence of similar communities on various topics and formed the Stack Exchange (SE) network.¹ The advancement of Information and communication technologies (ICTs) have enabled faster and easier creation and sharing of knowledge, but also the access to a large amount of data that allowed a detailed study of their emergence and evolution [6], as well as user roles [7], and patterns of their activity [8–10]. However, relatively little attention has been paid to the sustainability of SE communities. Most research focused on the activity and factors that influence the users' activity in these communities. Factors such as the need for experts and the quality of their contributions have been thoroughly investigated [11]. It was shown that the growth of communities and mechanisms that drive it might depend on the topic around which the community was created [12].

In this paper, we investigate the role of network structure and social trust dynamical user reputation in the sustainability of a knowledge-sharing community. Research on the sustainability of social groups shows that social interaction and their structure influence the dynamics and sustainability of social groups [13–16]. Due to large number of users and the smaller probability of repeated interactions dyadic trust between members may not play an essential role in the group dynamics of knowledge-sharing communities. However, it is known that the reputation of users, one of the proxies of trust in online communities, is the primary for them to become and maintain their productive member status [17–19].

With the proliferation of misinformed decisions, it is crucial to understand how to foster communities that promote collaborative knowledge exchange and understand how cooperative norms of trustworthy behavior emerge. The way people interact, specifically the structure of their interactions [20], and how inclusive and trustworthy the key members of the community can influence the sustainability of the knowledge-sharing communities. Although the topic and early adopters are essential in establishing a new SE community, they are not sufficient for sustainability. The current SE network has several examples of communities where the first instance of the community did not survive the SE evaluation process and was shut down, while the second attempt resulted in a sustainable community. Focusing on attempts to establish a community on the same or similar topic with a different outcome allows us to investigate the relevance of social network structure and social trust in the sustainability of knowledge-sharing communities. They are particularly relevant if we wish to understand why some communities established themselves in their second attempt. For those pairs of communities, the topic is the same, and all the initial

¹More information about Stack Overflow is available at: <https://stackoverflow.co/> and broad introduction to Stack Exchange (SE) network is available at: <https://stackexchange.com/tour>. Visit <https://area51.stackexchange.com/faq> for more details about closed and beta SE communities and the review process.



SE platform requirements were satisfied, but something else was crucial for community decay in the first attempt and its in the second.

Our methods and key results are summarised in a visual abstract in Fig. 1. In our main analysis, we analyze four pairs of SE communities and study the differences in the evolution of social structure and trust between closed and active communities. We have selected four topics from the STEM and humanities: astronomy, physics, economics, and literature. We focus on topics where we could find a matched pair of closed and active communities to control for the differences in topic popularity and, partially, community size. For this reason alone, we do not include Stack Overflow as the most popular community in our analysis. We analyze each pair's early stages of evolution and look at the differences between active and closed communities. Specifically, we map the interactions onto complex networks and examine how their properties evolve during the first 180 days of communities' existence. Using complex network theory [21] we quantify the structure of these networks and compare their evolution in active and closed communities on the same topic. We pay special attention to the core-periphery structure of these networks since it is one of the most prominent features of social networks [22]. We examine how core-periphery structure of active and closed communities evolve and analyze their difference. We show that active communities have a higher value of local normalized clustering and a more stable core membership. On average, the core of the sustainable communities has higher inner connectivity.

To study the evolution of social trust, we adapted the Dynamic Interaction Based Reputation Model (DIBRM) [23]. The model allows us to quantify the trust of each individual over time. We can quantify members' mean and total trust within the core and periphery and follow their evolution through time. The mean reputation of members is higher in sustainable communities than in closed ones, indicating higher levels of social trust. Furthermore, the mean reputation of core members of active communities is constantly above the mean reputation of core members in closed communities, indicating that the creation of trust in the early stages of a community's life may be crucial for its survival.

Our results show that social organization and social trust in the early phases of the life of a knowledge-sharing community play an essential role in its sustainability. Our analysis reveals differences in the evolution of these properties in communities on different topics.

The paper is organized as follows. In Sect. 2 we give a short overview of previous research. Section 3 describes the data and outlines some specific properties of each community. In Sect. 4 we describe the measures and models used for describing the local organization and measuring reputation. Section 5 shows our results. Finally, we discuss our results and selection of model parameters and time window, as well as its consequences in Sect. 6.

2 Previous research

The availability of data from the SE network led to detailed research on the different aspects of dynamics of knowledge sharing communities [6, 8–10], the roles of users [7], and their motivations to join and remain members of these communities [24–28]. The focus of the research in the previous decade was on the evolution of activity in SE communities and the different factors that influence this growth. Ahmed et al. [29] have investigated differences between technical and non-technical communities and showed that within the first four years, technical communities have a higher growth rate, more activity, and are more modular. The comparison of UX community in SE and Reddit [30] showed that the Reddit community grows faster, while SE becomes less diverse and active over time. Special attention was paid to the activities of individual users. In Ref. [31] authors argue that while the overall quality of the answers, measured in the answer score, decays over time, the quality of the answers of the individual user remains constant. This observation suggests that good answerers are *born* and not made within the community. Reputation is used as a proxy for the recognition of experts [32] by other members. However, contrary to common sense, the authors show that the presence of experts can reduce the activity of other members [32]. In [12] authors explore the role of self- and cross excitation in the temporal development of user activity. Differences between growing and declining communities and communities on STEM and humanities topics were explored. Their results show that the early stages of growing communities are characterized by the high cross-excitation of a small fraction of popular users. In contrast, later stages exhibit strong long-term self-excitation in general and cross-excitation by casual users. It was also shown that cross-excitation with power users is more important in the humanities than in STEM communities, where casual users have a more critical role.

A relatively small number of papers focus on the sustainability of SE communities. In Ref. [11], authors examine SE sites through an economic lens. They analyze the relationship between content production based on the number of participants and activities and show that an increase in the number of questions (input) increases the number of answers (output). In their works, Oliveira et al. [33] investigate activity practices and identify the tension between community spirit as proclaimed in SE guidance and individualistic values as in reputation measurement through focus groups and interviews.

Our assumption about the relevance of the structure of social networks in the sustainability of knowledge-sharing communities is supported by research on other social groups. Various factors influence the emergence [34, 35], the evolution, and the sustainability of the groups [13, 20, 36, 37]. The number of committed members [37] and the minimal level of interdependence between members [35] are important factors for the emergence of the

community. The levels of activity have an important role in the emergence and stability of social groups [34, 37], while social factors, such as the size of the group, number of social contacts, or social capital, influence their emergence and collapse [13–16].

Another important branch of research of interest in the sustainability of online communities is the topic of trust. While ICTs make it easier for individuals to establish and maintain social contacts and exchange information and goods, they are also exposed to new risks and vulnerabilities. Social trust relationships, based on positive or negative subjective expectations of another person's future behavior, play an important but largely unexplored role in managing those risks. Recent works show that the vital element of trust is the notion of vulnerability in social relations, and as negative expectations of a trustee's behavior most often imply damage or harm to the trustor, decisions about which users to trust in an online community become paramount [38–40].

In communities such as SE, individuals have three sources of information to rely on when deciding to trust someone in a specific context: (1) knowledge of previous interactions, (2) expectations about future interactions, and (3) indirect information gained through a broader social network. Suppose that the number of active users in such a community increases over a more extended period. In that case, the individuals have little or no history together, no direct interactions, and almost no memory of past interactions. In that case, the social network created by the community becomes a crucial source of information. Therefore, from a network perspective, trust can be the result of reputational concerns and can flow through indirect connections linking actors to one another [40, 41].

In that case, users rely on reputation as a public measure of the reliability of other users active within the same community. Reputation is often quantified based on the history of behavior valued or promoted by a set of community norms and, as such, represents a social resource within the community [42–44]. Since reputation is public information, it is also an incentive. Agents with high reputations are motivated to act trustworthy in the future in order to preserve their status in the community [41]. This idea is supported by psychological findings suggesting that trust is primarily motivated by effects produced by the act of trust itself, regardless of more rational or instrumental outcomes of trustworthy behavior [39].

In terms of modeling collective trust and reputation in online communities, knowledge about past behaviors can be implemented in a trust model in different ways. When estimating trust between agents in a social network, graph-based models focus on the topological information, position, and centrality of agents in a social network to estimate both dyadic and collective measures of social trust. On the other hand, interaction-based models, such as the dynamic reputation model implemented in this paper (DIBRM) [23] estimate trust or reputation based on the frequency and type of agent's interactions over time without taking into account the structure and topology of the interactions between different agents in a network.

3 Data

In our main analysis, we focus on pairs of closed and active SE communities matched by topic. Astronomy, Literature, and Economics are currently active communities. All three communities thrived the second time they were proposed. The first attempt to create communities on these topics resulted in website closure within a year. We add to the comparison the early days of the Physics community and compare its evolution with the closed

Theoretical Physics community. The topics of these communities are not identical, but it is safe to assume that there is a high overlap in user demographics and interests. For these reasons, we treat this pair in the same manner as others. Furthermore, to further solidify our results we have examined the early evolution of four additional communities: Mathematics, Mathematica, Startup Business, Startups. These communities are used to inspect the robustness of our main analysis by comparing main communities with others of similar size, user growth, and activity trends.

The SE data are publicly available and released at regular time intervals. We are primarily interested in the activity and interaction data, which means that we extract the following information for posts (questions and answers) and comments: (1) for each post or comment, we extract its unique ID, the time of its creation, and unique ID of its creator - user; (2) for every question, we extract information about IDs of all answers to that question and ID of the accepted answer; (3) for each post, we collect information about IDs of its related comments. The data contains information about the official SE reputation of each user but only as a single value measuring the final reputation of the user on a day when the data archive was released. Due to this significant shortcoming, we do not include this information in our analysis. In SE, users can give positive or negative votes to questions and answers and mark questions as favorites. However, the data is again provided as a final score recorded at the release. Since this does not allow us to analyze the evolution of scores, we omit this data from our analysis.

All SE communities follow the same path from their creation until they are considered mature enough or closed. In a *Definition* phase, a small number of SE users start by designing a community by proposing hypothetical questions about a certain topic. A successful *Definition* phase is followed by a *Commitment* phase. In this phase, interested users commit to the community to make it more active. The *Beta* phase, which follows after the *Commitment* phase, is the most important. It consists of two steps: a three-week private beta phase, where only committed users may ask/answer/comment questions, and a public beta phase when other members are allowed to join the community. The duration of the public beta phase is not limited. Depending on this analysis, there are three possible outcomes: (1) the community is considered successful and it graduates; (2) the community is active but needs more work to graduate, which means that the public beta phase continues; (3) the community dies and the site is closed. The community evaluation/review process is guided by simple metrics: the average number of questions per day, average number of answers per question, percentage of answered questions, total number of users and number of avid users, and average number of visits per day. However, it should be noted that process is not straightforward and that decision criteria have substantially changed in previous years and sometimes exceptions are made for specific communities.²

We study how the social network properties of these social communities and the social trust created among their members evolve during the first 180 days. The first 90 days are recognized as the minimal time a newly established community should spend in the beta phase. We investigate a period that is twice as long since closed communities were active between 180 and 210 days. Given that differences in the first few months of the life of the

²For example, in 2022 59 websites graduated according to new criteria established in 2019 (which excluded questions per day metric), but as explained in the announcement (<https://meta.stackexchange.com/questions/374096/congratulations-to-the-59-sites-that-just-left-beta>) exception was made for the AI community which graduated although it didn't meet the criteria that minimum 70% questions have at least one upvoted answer.

Table 1 Community overview for first 180 days according to SE evaluation criteria

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

online community can help predict its survival and evolution [45], we focus on the early evolution of SE sites.

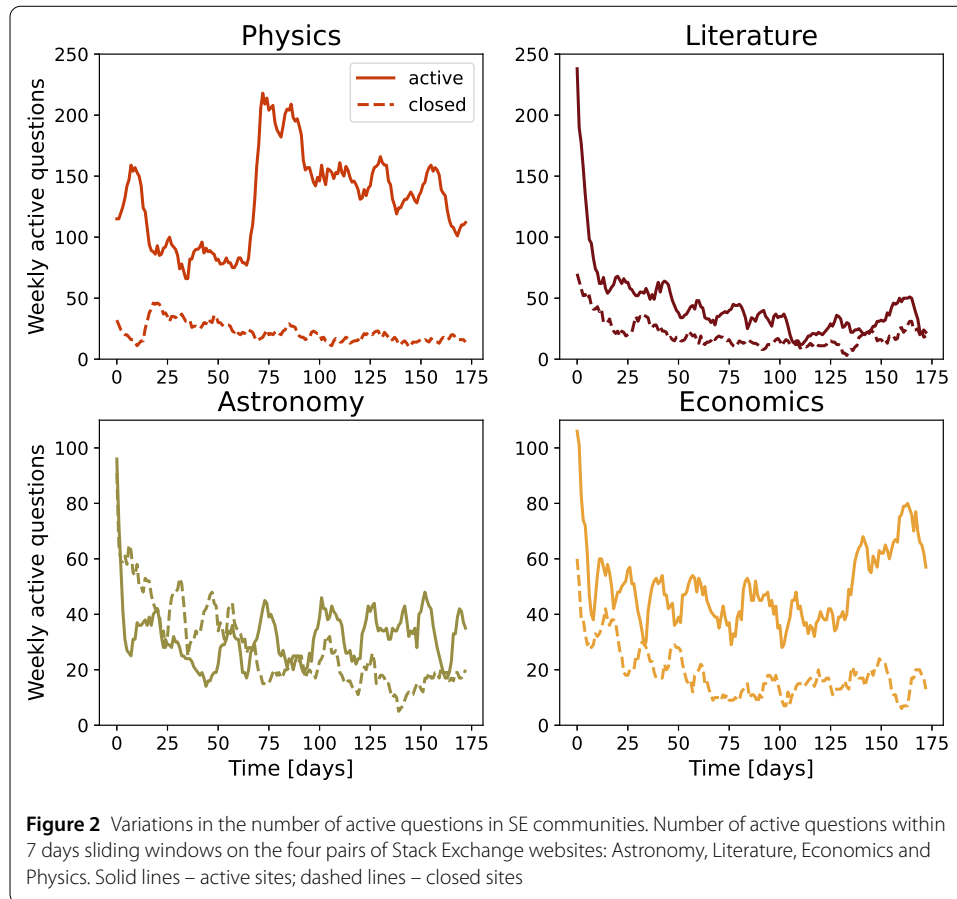
Although the official review of SE communities in the beta phase is mostly based on simple activity indicators such as the number of questions or ratio of answers to questions,³ these simple metrics do not provide enough information to differentiate between closed communities and those that have been proven to be sustainable in the long term. This may explain why the official guidelines for SE community review have changed and have been applied inconsistently.

Table 1 shows the values of some of these measures at 180 days point for considered communities. Although the Physics community had better metrics than Theoretical Physics and other considered communities, we see that these differences are not as apparent if we compare the remaining three pairs of communities. For instance, some of the parameters for the closed Astronomy community, for example, the percentage of answered questions and answer ratio, were better than for the community that is still active.

Another simple indicator can be the time series of active questions for the 7 days shown in Fig. 2. The question is considered active if it had at least one activity, posted answer, or comment, during the previous 7 days. The four pairs of compared communities show that active communities have a higher number of active questions after 180 days. Although this difference is evident for the Physics and Economics community, Fig. 2 shows that its value is smaller for Astronomy and Literature. Furthermore, in the case of Astronomy, the closed community had a higher number of active questions in the first 75 days.

The values of the measures shown in Tables 1 and A1 in Additional file 1, and Fig. 2 suggest that these simple measures are not good indicators of long-term sustainability. Therefore, we need a deeper understanding of the structure and dynamics of the community to understand the factors behind its sustainability. All communities must start with the same number of interesting questions, the same number of committed users, and satisfy the same thresholds to enter the public beta phase. These basic aggregated statistics are not enough to differentiate between active and closed communities. Hence, other factors determine the sustainability of communities. We investigate the role of social interaction structure and the dynamics of collective trust in the sustainability of SE communities.

³<https://stackoverflow.blog/2011/07/27/does-this-site-have-a-chance-of-succeeding/>



4 Method

We are interested in the position of trustworthy members in SE communities and how active and closed communities differ regarding this factor. First, we map the interaction data onto networks and analyze their properties and how they evolve during the first 180 days. Furthermore, we use the dynamical reputation model to estimate the trustworthiness of each member of the community and the dynamics of collective trust by studying the evolution of the mean value of reputation in the community. The entire analysis was done in Python, and the entire code for reproducing the results and figures is publicly available in an online repository.⁴

4.1 Network mapping

We treat all user interactions, answering questions, posting questions or comments, and accepting answers equally. We construct a network of users where the link between two nodes, users i and j , exists if i answers or comments on the question posted by j and vice versa, or i comments on the answer posted by j and vice versa, i accepts the answer posted by user j . We do not consider the direction or frequency of the interaction between users i and j ; thus, the obtained networks are unweighted and undirected.

We create a network snapshot $G(t, t + \tau)$ at the time t for the time window length τ . Two users (i, j) are connected in a network snapshot $G(t, t + \tau)$ if they have had at least one

⁴<https://github.com/ana-vranic/Stack-Exchange-communities>

interaction during the time $[t, t + \tau]$. Our first network accounts for interaction within the first 30 days $G[0, 30)$, and we slide the interaction window by one day and finish with $G[149, 179)$ network. This way, we create 150 interaction networks for each community. By sliding the time window by one day, we create two consecutive networks that overlap significantly. In this way, we can capture subtle structural changes resulting from daily added/removed interactions. We calculate the different structural properties of these networks and analyze how they change over 180 days.

4.2 Clustering

There are many local and global measures of network properties [21]. These measures are not independent. However, it was shown that the degree distribution, degree-degree correlations, and clustering coefficient are sufficient to fully describe most complex networks, including social networks [46]. Furthermore, research on the dynamics of social group growth shows that links between persons' friends who are members of a social group increase the probability that that person will join that social group [47]. Successful social diffusion typically occurs in networks with a high value of the clustering coefficient [48]. These results suggest that higher local cohesion should be a characteristic of sustainable communities.

The clustering coefficient of a node quantifies the average connectivity between its neighbors and the cohesion of its neighborhood [21]. It is a probability that two neighbours of a node i are also neighbours, and is calculated using the following formula:

$$c_i = \frac{e_i}{\frac{1}{2}k_i(k_i - 1)} \quad (1)$$

Here e_i is the number of links between the neighbours of the node i , while $\frac{1}{2}k_i(k_i - 1)$ is the maximum possible number of links determined by the degree of the node k_i . The clustering coefficient of the network C is the value of the clustering averaged over all nodes. We investigate how the clustering coefficient in an SE community changes over time by calculating its value for all network snapshots. We normalize the clustering coefficients with the value of expected clustering for the random Erdos-Renyi network with the same number of nodes N and links L : $c_{er} = p = \frac{2L}{(N(N-1))}$ [21, 49]. We compare normalized clustering coefficient for active and closed communities on the same topic to better understand the evolution of cohesion of these communities.

4.3 Core-periphery structure

Real networks, including social networks, have a distinct mesoscopic structure [22, 50]. The mesoscopic structure is manifested either through the community structure or the core-periphery structure. Networks with a community structure consist of a certain number of groups of nodes that are densely connected, with sparse connections between groups. Networks with core-periphery structures consist of two groups of nodes, with higher edge density within one group, core, and between groups. However, low edge density in the second group, periphery [22]. Research on user interaction dynamics in SE communities shows that there is a small group of highly active members who have frequent interactions with casual or low active members [8, 12]. These results indicate that we should expect a core-periphery structure in SE communities. The classification of nodes

into one of these two groups provides information on their functional and dynamic roles in the network.

To investigate the core-periphery structure of SE communities and how it evolves over time, we analyze the core-periphery structure of every network snapshot. For this purpose, we use the Stochastic Block Model (SBM) adapted for the inference of the core-periphery of the network structure [22].

SBM is a model where each node belongs to one group in the given network G . For the core-periphery structure, the number of blocks is two. Thus, the elements of the vector θ_i are 1 if the node i belongs to the core or 2 for the periphery. The block connectivity matrix $\{\mathbf{p}\}_{2 \times 2}$ specifies the probability p_{rs} that nodes from group r are connected to nodes in group s , where $r, s \in \{1, 2\}$.

The SBM model seeks the most probable model that can reproduce a given network G . The probability of having model parameters θ, \mathbf{p} given network G is proportional to the likelihood of generating network G , $P(G|\theta, \mathbf{p})$, prior on SBM matrix $P(\mathbf{p})$ and prior on block assignments $P(\theta)$:

$$P(\theta, \mathbf{p}|G) = P(G|\mathbf{p}, \theta)P(\mathbf{p})P(\theta), \quad (2)$$

The likelihood of generating a network G is defined as:

$$P(G|\theta, \mathbf{p}) = \prod_{i < j} p_{r_i s_j}^{A_{ij}} (1 - p_{r_i s_j})^{1 - A_{ij}}, \quad (3)$$

where the adjacency matrix element A_{ij} is equal to 1 whenever nodes i and j are connected and it is 0 otherwise.

Prior on \mathbf{p} is the uniform distribution over all block matrices whose elements satisfy the constraint for the core-periphery structure $0 < p_{22} < p_{12} < p_{11} < 1$. Prior on θ consists of three parts: the probability of having 2 blocks; given the number of blocks, probability $P(n|2)$ of having groups of sizes $\{n_1, n_2\}$ and probability $P(\theta|n)$ of having particular assignments of nodes to blocks.

To fit the model, we follow the procedure set by the authors of Ref. [22] and use the Metropolis-within-Gibbs algorithm. For each 30 days snapshot network, we run 50 iterations and choose the model parameters θ and \mathbf{p} according to the minimum description length (MDL). MDL does not change much among inferred core-periphery structures, see Fig. A1 in Additional file 1, while looking into the Adjusted Rand Index (ARI), we can notice that difference exists. Still, the ARI between pair-wise compared partitions is significant (ARI > 0.9), indicating the stability of the inferred structures. The definition and detailed descriptions of MDL and ARI are given in the Additional file 1.

4.4 Dynamic 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 [43]. First, the dynamics of trust are asymmetric, meaning that trust is easier to lose than to gain. As part of asymmetric dynamics, to make trust easier to lose, the trust metric has to be sensitive to new experiences, recent activity, or the absence of the user's activity while still maintaining the non-trivial influence of old behavior. The impact of

new experiences must be independent of the total number of recorded or accumulated past interactions, making high levels of trust easy to lose. Finally, the trust metric must detect and penalize behavior that deviates from community norms.

We estimate the dynamic reputation of SE users using the Dynamic Interaction Based Reputation Model (DIBRM) [23]. This model is based on the idea of dynamic reputation, which means that the reputation of users within the community changes continuously over 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 in users' reputations is found through bursts of activity followed by a short period of inactivity.

Our model implementation does not distinguish between positive and negative interactions in SE communities. Therefore, we treat any interaction in the community, posting a question, answer, or comment, as a potentially valuable contribution. The evaluation criteria for SE websites that go through beta testing described in Additional file 1 do not distinguish between positive and negative interactions. The percentage of negative interactions in the communities we investigated was below 5%, see Table A2 in Additional file 1. Filtering positive interactions would also require filtering out comments because the community does not rate them. That would eliminate a large portion of direct interactions between community users, which is essential for estimating their reputation. The only negative aspect of behavior in our model is the absence of valuable contributions - the user's inactivity. This behavior can be seen as a deviation from community norms as we look at new communities in the early stages of development, where constant contributions are crucial to community growth and survival.

In DIBRM, the reputation value for each user of the community is estimated by combining two different factors: (1) *reputation growth* - the cumulative factor that represents the importance of users' activities; (2) *reputation decay* - the forgetting factor that represents the continuous decrease in reputation due to inactivity. In the case of SE communities, the forgetting factor has a literal meaning, as we can assume that active users forget users' past contributions as their attention is captured by more recent content.

In the bottom left part of Fig. 1 we see an example of reputation dynamics for a single user. There are bursts of reputation growth after multiple interactions are recorded, like in the case of two interactions in a single day recorded between days 25 and 50, followed by a period of inactivity which leads to reputation decay. In this case, the decay is interrupted by a single recorded activity before the 75th day, but then an even longer inactivity period ensued, leading to a decay that reduced the reputation of the user nearly to 0 before the 100th day. Two contrasting examples of real user reputation are explained in the Additional file 1 (Fig. A2).

Reputation dynamics revolves around the varying influence of past and recent behavior. Thus, 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} . The cumulative factor I_{c_n} then captures the additive effect of successive recent interactions. In Fig. 1 we see this cumulative effect with two consecutive interactions (gray vertical lines) after day 150 which sudden jump in reputation previously reduced to zero. The reputational contribution I_n of the most recent interaction n of any given user is estimated in the following

way:

$$I_n = I_{b_n} + I_{c_n} = I_{b_n} \left(1 + \alpha \left(1 - \frac{1}{S_n + 1} \right) \right). \quad (4)$$

Here, α is the weight of the cumulative part, and S_n is the number of sequential activities. If there is no interaction at t_n , this part of interactions has a value of 0. An essential property of this component of dynamic reputation is the notion of sequential activities. Two subsequent interactions by a user are considered sequential if the time between these two activities is less than or equal to the time parameter t_a that represents the time window of interaction. This time window represents the maximum time spent by the user to make a meaningful contribution, post a question or answer, or leave a comment,

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a}. \quad (5)$$

If $\Delta_n < 1$, the number of sequential activities S_n will increase by one, which means that the user continues to communicate frequently. However, large values Δ_n significantly increase the effect of the forgetting factor. This factor plays a vital role in updating the total dynamic reputation of a user at each time step, after every recorded interaction:

$$T_n = T_{n-1} \beta^{\Delta_n} + I_n. \quad (6)$$

Here, β is the forgetting factor. In our model implementation, the trust is updated each day for every user regardless 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 β lead to faster trust decay, as shown in Fig. A2 in the Additional file 1. In Fig. 1 we observe this long-tailed reputation loss when the user has more than 25 inactive days between days 120 and 150, reducing the reputation almost to 0.

For this work, we select the following values of these parameters: (1) we set the basic reputation contribution $I_{b_n} = 1$, which means that each activity contributes 1 to the dynamical reputation; (2) for the cumulative factor α we choose the value 2 and place higher weight on recent successive interactions; (3) forgetting factor β we select the value 0.96; (4) the value of $t_a = 2$. By setting $\alpha > 1$ we enable faster growth of reputation due to a large number of subsequent interactions; see Fig. A2 in Additional file 1. Furthermore, by setting the value of $\beta < 1.0$, we increase the penalty for long inactivity periods; see Fig. A2 in Additional file 1. We discuss the selection of model parameters and their consequences in detail below. The selected values of parameters are used to measure the dynamical reputation of users in all four pair SE communities. Given these parameter values, the minimal reputation of the user immediately after having 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 window. Users with a reputation below the value 1 are considered inactive and *invisible* in the community; that is, their past contributions at that time are unlikely to impact other users.

4.4.1 The choice of model parameters

In this work, we used snapshots of the network of 30 days. This period corresponds to the average month, and it is common in the analyses of the structure and dynamics of social networks [51–53]. Still, there is no well-specified procedure to choose the time window. Previous studies have shown that if τ is small, subnetworks become sparse, while for too large sliding windows, some important structural changes cannot be observed [52, 54]. Thus, we have analysed how the time window choice influences our results. Figure A11 in Additional file 1 shows how considered network properties and dynamical reputation depend on the time window size for active and closed communities in case of Astronomy communities. We observe that fluctuations of all measures are more pronounced for a time window of 10 days than for 30 and 60 days. However, we find that while the structural properties of networks evolve at different rates over varied time windows, the trends remain very similar. The qualitative difference observed between closed and active communities is independent of the time window size, especially when comparing the 30 and 60 day windows. The 30-day time window ensures enough 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.

The initial purpose of DIBRM was to replicate the dynamics of the official SE reputation metric [23, 55]. In previous studies [55] the official SE reputation is obtained with $t_a = 2$, $\alpha = 1.4$, $\beta = 1$. This configuration of model parameters implies that there is no reputation decay and points toward the fact that the official SE reputation is hard to lose. Our application is oriented towards estimating a reputation metric which takes into account the fundamental properties of social trust, i.e. reputation decreases with members' inactivity, so we opted for a different set of parameter values.

For the 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 the initial contribution of +1 to the 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 the 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.

The combination of parameters α and β can significantly influence the dynamic of the single user reputation, as shown in Fig A2. We show that higher values for parameter $\alpha = 2$, highlight the burst of user activity and frequent interaction. On the other hand, the parameter beta 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 parameter β value, so we include forgetting due to inactivity but do not penalize it too much. In Fig. A2, we show how different values of parameter β influence the time needed for a user's reputation to fall on value $I_n = 1$ due to the user's inactivity and value of dynamical reputation at the moment of the last activity. The higher the value of the parameter β and the initial dynamical reputation of the users, the longer it takes for the user's reputation to fall to the baseline value. For parameters $\beta = 0.9$ and $I_n = 5$, the user's reputation drops to 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 the parameter β , the time it takes for I_n to drop to 1 becomes longer and that the initial value of the reputation becomes less important.

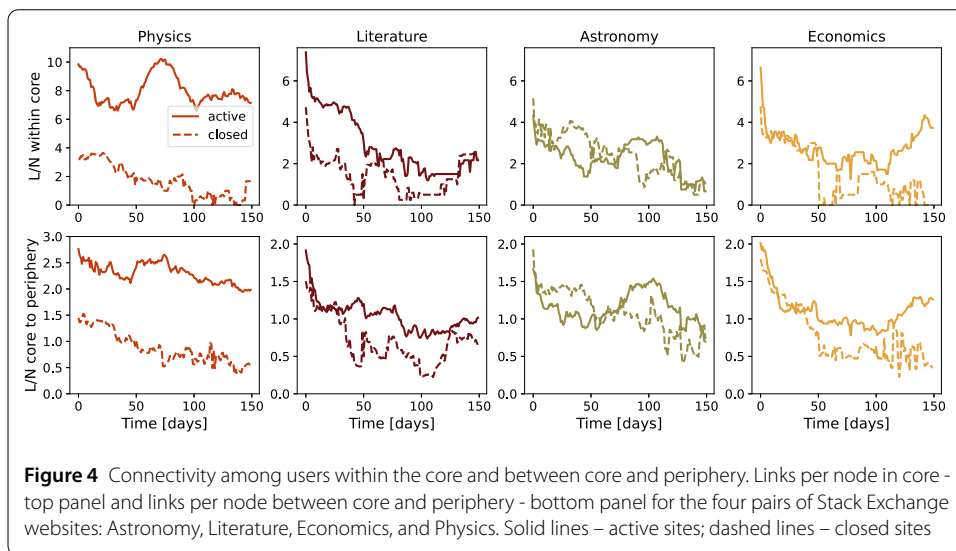
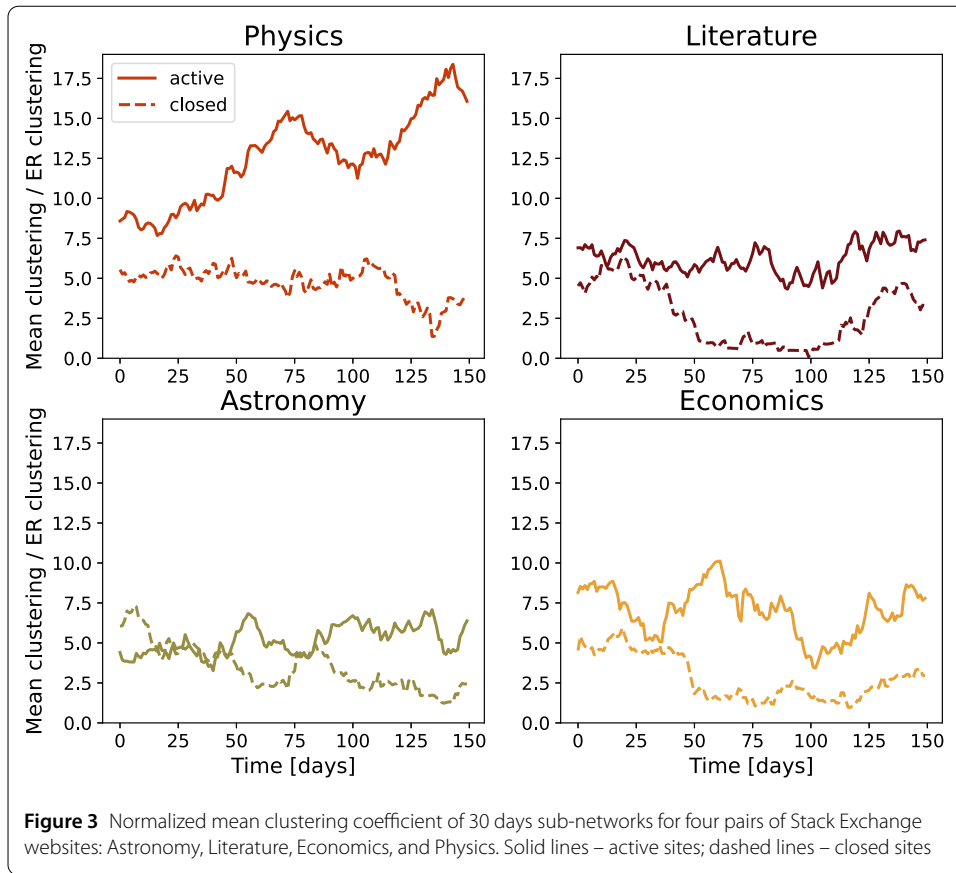
We estimated the difference between the number of users who had at least one activity in the 30-day window and the number of users with a reputation greater than 1 during the same period for different parameter β values. We calculated the root mean square error (RMSE) between the time series of the number of active users for $\tau = 30$ and different values of β parameters; see Fig. A12 in Additional file 1. The minimal difference between these two variables is for β between 0.94 and 0.96 for both active and closed communities. Since we want to compare communities, we select $\beta = 0.96$. Our analysis reveals that the reputational decay parameter β set at 0.96 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; see Fig. A13 in Additional file 1. Furthermore, we examine and compare the trends of two types of time series: (1) time series of active users, according to dynamical reputation; (2) time series of permanent users, users who were active in a given sliding window and continued to be active in the next one. Figure A14 in Additional file 1 shows that while the absolute number of users differs in these time series, they follow similar trends for all communities.

5 Results

5.1 Clustering and core-periphery structure of knowledge-sharing networks

We first analyze the structural properties of SE communities and examine the difference between active and closed ones. We calculate the normalized mean clustering coefficient for 30-day window networks and examine how it changes over time. Figure 3 shows the evolution of the normalized mean clustering coefficient for the eight communities. All communities that are still active are clustered, with the value of normalized clustering coefficient above 5, with Physics, the only launched community, having the highest value of normalized clustering coefficient during the first 180 days. During the larger part of the observed period, an active community's normalized clustering coefficient is higher than the normalized clustering coefficient of its closed pair. For pairs where active communities are still in the beta phase, some of closed communities have a higher value of the normalized clustering coefficient in the first 50 days. After this period, active communities have higher values of the normalized clustering coefficient. These results suggest that all communities have relatively high local cohesiveness compared to random graphs, however, the value of normalized clustering below the value 5 in the later phase of community life may indicate its decline.

Furthermore, we examine the core-periphery structure of these communities and their evolution. Specifically, we are interested in the evolution of connectivity in the core. Figure 4 shows the change in the number of links between nodes, averaged on the core nodes, $\frac{L_c}{N_c}$ over time. $\frac{2L_c}{N_c}$ is the average degree of the node in the core and, thus, $\frac{L_c}{N_c}$ is the half of the average degree. Again, the Physics community has a much higher value of this quantity than Theoretical Physics during the observed period, indicating higher connectivity between core members. Higher connectivity between core members in the active community is also characteristic of Literature. However, 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 prominent for Economics and Astronomy, see Fig. 4. 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 is twice as large as in the closed community, and the difference grows at



the end of the observation period. The connectivity in the core of the closed Astronomy community is higher than the connectivity in the core of the active community during the first 50 days. However, as time progresses, this difference changes in favor of the active community, while this difference disappears at the end of the observation period.

The difference between active and closed communities is observed compared to the average number of core-periphery edges per network node. The connectivity between core

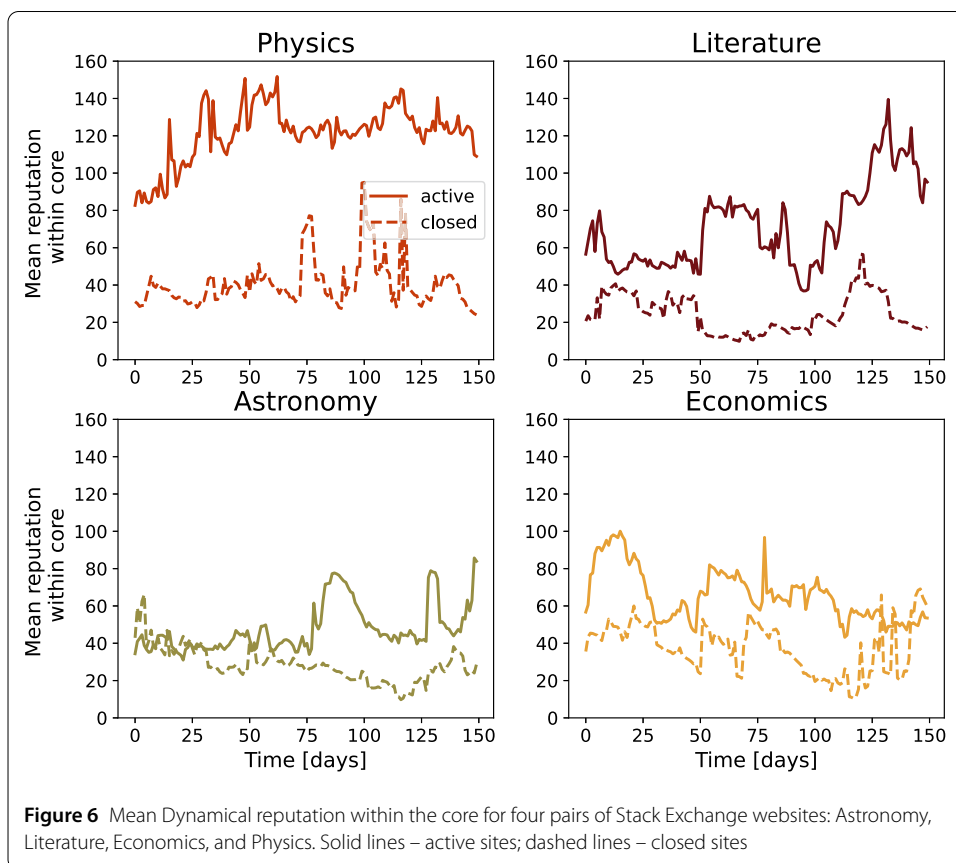
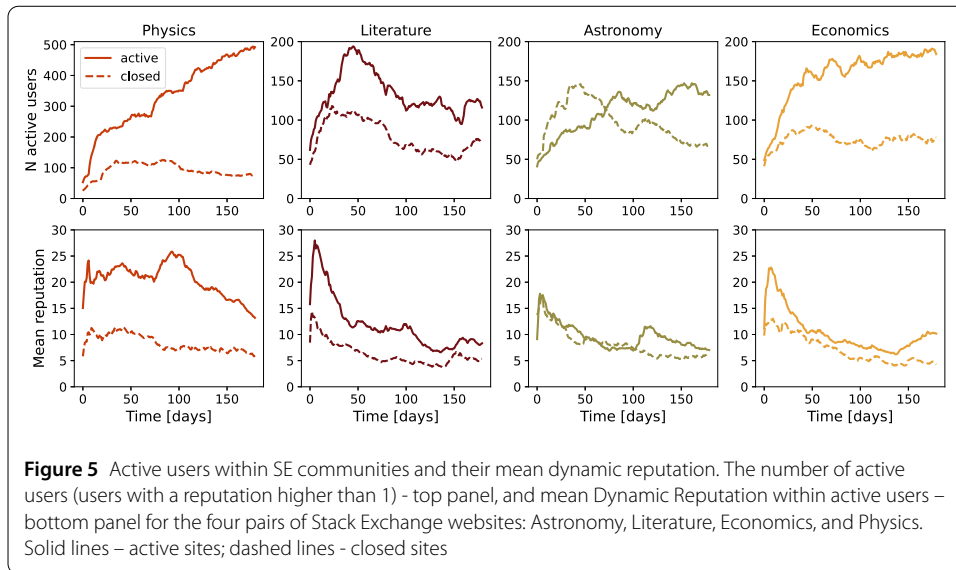
and periphery is higher for the active communities than for the closed ones, see Fig. 4, which is very obvious if we compare Physics and Theoretical Physics communities. Moreover, the Physics community has the highest connectivity compared to all other communities. Active Literature and Economics communities have the same core-periphery connectivity as their closed counterpart. The core of the active Astronomy community has weaker connections with the periphery than the closed community during the first 50 days, see Fig. 4.

Our motivation to examine the core-periphery structure comes from reference [12]. The authors have selected 10% of the most active users and examined their mutual connectivity and connectivity with the remaining users. The split of 10% to 90% users according to their activity may appear arbitrary. The core-periphery provides a more consistent network division based on its structure. However, the connectivity patterns between popular-popular and popular-casual users, shown in Fig. A3 in Additional file 1, are similar to one observed for core-periphery in Fig. 4.

On average, the cores of active communities have a higher number of nodes than closed communities. However, the size of the core relative to the size of the network is similar for active and closed communities (Fig. A4 in Additional file 1). The size of the core fluctuates over time for active and closed communities. The core membership also changes over 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 at the moment t_i and t_j . Figure A5 in Additional file 1 shows the value of the Jaccard index between any pair of the 150 subnetworks. The highest value of the Jaccard index is around the diagonal and has a value close to 1. The compared subnetworks are for consecutive days and have a similar structure. The value of the Jaccard index decreases with the number of days between two subnetworks $|t_i - t_j|$ faster in closed communities; see Fig. A6 in Additional file 1. This difference is the most prominent for the Literature communities, while this difference is practically non-existent for Astronomy. The relatively high value of overlap between cores of distant subnetworks for active communities further confirms that the core is more stable in these communities than in their closed counterparts.

5.2 Dynamic reputation of users within the network of interactions

To explore the differences between active and closed communities, we focus on dynamical reputation, our proxy for collective trust in these communities. The number of active users (top panel) and the mean user reputation (bottom panel) for different SE communities are shown in Fig. 5. Except in the case of Astronomy, closed communities generated less engaged users from the start and the number of active users saturated at lower values. In the case of Astronomy, the closed community started with a faster-increasing number of active users. However, within the first two months, their number dropped, while the second time around, the community started slower but kept engaging more users. Only in the still active Physics community is the number of active users an increasing function over the whole 180 day period we have observed. Panels in the bottom show mean reputation among active users, and we see that most of the time, it was higher in the still active communities than in the closed ones. The Physics community kept these mean values more stable at higher levels, whereas in other communities, we note that the initial high mean reputation decays faster. Astronomy is an exciting exception again, where we see a second sudden increase in mean user reputation, which signals an increase in user activity.



In addition, we investigate whether and how the core-periphery structure is related to collective trust in the network. Figure 6 shows the mean dynamical reputation in the core of active and closed communities and its evolution during the observation period. There are apparent differences between active and closed communities regarding dynamical reputation. The mean dynamical reputation of core users is always higher in active commu-

nities than in closed. The most significant difference is observed between the Physics and Theoretical Physics communities. 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 a higher mean dynamical reputation, especially in the later phase of the observation period. The only difference in the pattern is observed for Astronomy communities at the early stage of their life. The closed community has a higher value of dynamic reputation than the active community. This observation is in line with similar patterns in the evolution of mean clustering, core-periphery structure, and mean reputation.

By definition, the core consists of very active individuals. Thus we expect a higher total dynamical reputation of users in the core than the total reputation of users belonging to subnetworks periphery. Figure A7 in Additional file 1 shows the ratio between the total reputation of the core and periphery for closed and active communities and their evolution. The ratio between the total reputation of core and periphery in Physics is always higher than in the Theoretical Physics community. A similar pattern can be observed for Literature communities, although the difference is not as prominent as in the case of Physics. The ratio of total dynamical reputation between core and periphery was higher in the closed Economics community during the early days of its existence. However, this ratio becomes higher for active communities in the later stage of their lives. Communities around the astronomy topic deviate from this pattern, which shows the specificity of these two communities.

To complete the description of the evolution of dynamic reputation, we examine the evolution of the Gini index of dynamical reputation among the active members of SE sites, shown in Fig. A8 in Additional file 1. Both closed and active communities have high values of the Gini index, indicating that the dynamic reputation is distributed unequally among users. Notably, all communities have the highest Gini index at the start, signaling that the inequality in users' activity at the start, and thus their dynamic reputation is the highest. After this initial peak, the Gini index decreases, but it persists at higher levels in communities that are still active than in the closed ones, except in the case of the Astronomy community. In this case, the active community had a higher Gini index until just before the observation period, when the Gini coefficient increased in the closed community.

Figure A9 in Additional file 1 shows the evolution of the assortativity coefficient for users' dynamical reputation. The observed networks are disassortative during the most significant part of 180 days period. Users with high dynamical reputations tend to connect with users with a low value of dynamical reputation in all eight communities. We also compare the degree and betweenness centrality of the users and their dynamical reputation by calculating the correlation coefficient between these measures for each sliding window, see Fig. A10 and detailed explanation in Additional file 1. The correlation between these centrality measures and dynamical reputation is very high. In active communities on physics, economics, and literature topics, the correlation between centrality measures and users' reputation is exceptionally high, above 0.85, and does not fluctuate much during the observation period. There is a clear difference between active and closed communities for these three pairs. The Astronomy pair deviates from this pattern for the first 100 days. After this period, the pattern is similar to one observed for the other three pairs of communities. The results reveal that degree and betweenness centrality are correlated more with a reputation in active than in closed communities.

6 Discussion and conclusions

In this work, we have explored whether the structure and dynamics of social interactions determine the sustainability of knowledge-sharing communities. We have adopted a model of dynamical reputation to measure the collective trust of members and analyzed its dynamics. For this purpose, we use the data from the SE platform of knowledge-sharing communities where members ask and answer questions on focused topics. We selected four pairs of active and closed communities on the same or similar topic. Specifically, two topics are from the STEM field, physics, and astronomy, and two are from social sciences and humanities, economics and literature.

We have examined the evolution of the normalized average clustering coefficient in closed and active SE communities. Our results show that active communities have significantly higher values of clustering coefficient compared to ER graphs of the same size in the later phase of community life than closed communities. In the early phase of communities' lives, the clear difference between active and closed communities is observed only for the physics topic; see Fig. 3. The high value of the normalized clustering coefficient observed for the active Physics community suggests that communities with high local cohesiveness are sustainable and mature faster than others.

The core in active communities is more strongly connected with the periphery than in closed communities, indicating that active members engage more often with occasionally active members; see Fig. 4. These results suggest that active communities are more inclusive than closed ones. Furthermore, our analysis shows that average connectivity between core members is not as crucial to community sustainability as expected. Although active Physics and Economics communities exhibit much higher connectivity in the core than their closed counterparts, this is not true for communities focused on astronomy and literature. However, our results show that a member's lifetime in the core is longer for active communities, indicating a more stable core in active communities.

Analysis of the evolution of the core-periphery and its connectivity patterns suggests a higher trust between active and sporadically active members. To further explore this, we have adapted the dynamical reputation model [23], which allowed us to follow the evolution of trust of each member.

The total dynamical reputation of core members during their first 180 days was higher for active communities than for their closed counterparts. While relative core size is less than 40%, Fig. A4 in Additional file 1, the ratio between the total reputation of nodes in the core and ones in the periphery is consistently above 0.5, indicating that the average reputation of members in the core is higher than the reputation of the node in the periphery. The ratio between the total reputation of core and periphery nodes has a higher value in the active community of Physics, Literature, and Economics. For most of the 180 days, this ratio has a value higher than one. The Astronomy communities are outliers, but the core members have a higher total reputation than members on the periphery, even for these two communities. Our results imply that the most trusted members in the community are the core members, who also generate more trust in active communities. They have a higher reputation generated through interactions with both core and nodes in the periphery, see Fig. 6. Furthermore, the overall levels of trust are higher in active communities, which is reflected in the fact that the mean user reputation is higher in these communities; see Fig. 5.

The choice of the topics and selection of SE communities of a various number of users, question, answer and comments, see Table A1 in the Additional file 1, guarantees, up to a certain extent, the generality of our results. However, there are certain limitations to the generalizability of our findings. While SE communities provide very detailed data that enable the study of the structure and dynamics of knowledge-sharing communities, we must not ignore the fact that they have some properties that make them specific.

SE communities are about specific topics; they mostly bring together people who are passionate about or are experts in a specific field. These communities attract people from the general population. Since we were interested in excluding the factor of the topic in our research, we studied and compared active and closed communities on the same topic. In the SE network, these pairs of communities are pretty rare, which has substantially limited our sample size, leaving the possibility for the occurrence of outliers that do not follow our general conclusions.

To further solidify our results, we have examined the early evolution of four additional communities: Mathematics, Mathematica, Startup Business, and Startups. Mathematics and Mathematica communities graduated early in the process, while both communities on startup topics were closed after spending some time in the public beta phase. Figures A15 and A16 in the Additional file 1 show that both communities on the subject of mathematics exhibit a similar evolutionary path as the Physics community. They have a high mean reputation, stable and relatively large cores with high average trustworthiness of core members, see Fig. A15 in Additional file 1. While the numbers of active users in these two communities and the Physics community differ, we see that this does not influence the average reputation of users or the size of the core. This is even more evident if we compare the Physics community with the closed Startup Business community. We see from Fig. A16 in Additional file 1 that the number of active users grows much faster for this community than for Physics. However, the average reputation in the community is comparable with the ones that were eventually closed, Theoretical Physics and Startups. Furthermore, the core size is comparable with the core of Physics, but the average trustworthiness of core members is similar to one for closed communities. These results demonstrate that even the communities with high early activity and a number of active users will not become sustainable if they do not develop a core of trustworthy members. Startups community has a behavior very similar to Theoretical Physics community. The comparison between two startup communities, shows that despite their difference in the activity levels these communities have similar evolution path during the first 180 days.

We have also decided to map interactions to networks so that the resulting network is unweighted and undirected. We use unweighted edges for a finer distinction between the structure and community dynamics. The number of repeated user interactions is captured with dynamic reputation, while the edges carry only structural information without the number of repeated interactions. Furthermore, as we map interactions to networks using sliding windows, the repeated presence of an edge throughout different windows gives us partial information about the durability and the frequency of the dyadic relationship. Similarly, we opted against directed weights as we are not interested in diffusion or flow of information and undirected edges represent a more parsimonious view of the community structure. However, these choices did have consequences in the choice of core-periphery detection method, and it is possible that with different network mapping, other methods would prove more suitable.

Finally, there are many ways to measure collective trust and reputation in online social communities. We have selected the dynamical reputation model because it was developed to measure reputation in SE communities. Furthermore, the model allowed us to study the evolution of trust in communities. However, the model requires fine-tuning of its parameters and does not distinguish positive from negative interactions. We have selected our parameters to replicate the activity of the SE communities in the time window of $\tau = 30$ days. Our analysis shows that while the choice of the sliding window, τ , may seem arbitrary, the different values do not influence the general conclusions; see Fig. A11 in Additional file 1. The interactions in SE communities are mostly not emotional, and thus, the model is suitable for measuring collective trust in these communities. However, the interaction in other knowledge-sharing communities can be much more emotional, and therefore the dynamical reputation model needs to be adapted to measure reputation in these communities.

Our results show that the trustworthiness of core members thus represents one of the essential parameters for determining community sustainability. Sustainable communities have a core of trustworthy members. The core of sustainable communities is more densely connected, and its connectivity with the periphery is more significant than in closed communities. The observed feature is especially prominent in the Physics community, which is the only active community considered to be mature. As we stated, active communities on topics of astronomy, economics and literature were in the beta phase. However, since December 2021,⁵ these communities graduated. The core of sustainable communities exhibits higher degrees of stability during their first 180 days. Sustainable communities have higher local cohesiveness, which is reflected in the relatively high value of the normalized clustering coefficient. Our results show that these conclusions hold for both STEM and humanities topics. However, we do not observe apparent differences between active and closed Astronomy communities for some quantities. In the case of Astronomy and sometimes Economics, we find that closed communities had higher normalized clustering coefficients and higher core-core and core-periphery connectivity during the early phase of community life. These observations suggest that the properties of the network during the early phase of the community's existence may lead to wrong conclusions about its sustainability. Our results also imply that information about community sustainability is hidden in the evolution of different network and trust properties.

Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1140/epjds/s13688-023-00381-x>.

Additional file 1. The file contains all additional figures, tables and descriptions regarding the analysis performed in the manuscript. The file is in pdf format. (PDF 3.6 MB)

Acknowledgements

Numerical simulations were run on the PARADOX-IV supercomputing facility at the Scientific Computing Laboratory, National Center of Excellence for the Study of Complex Systems, Institute of Physics Belgrade.

Funding

AA, AV and MMD acknowledge funding provided by the Institute of Physics Belgrade, through the grant by the Ministry of Education, Science, and Technological Development of the Republic of Serbia.

⁵<https://stackoverflow.blog/2021/12/16/congratulations-are-in-order-these-sites-are-leaving-beta/>

Abbreviations

ARI, Adjusted Rand Index; DIBRM, Dynamic Interaction Based Reputation Model; ICT, Information and communication technologies; MDL, Minimum Description Length; RMSE, Root mean square error; SBM, Stochastic Block Model; SE, Stack Exchange.

Availability of data and materials

The Stack Exchange data can be downloaded from Stack Exchange Data Dump, <https://archive.org/details/stackexchange>. Area 51 Stack Exchange communities can be downloaded from <https://area51.stackexchange.com/>. The source code and the datasets generated and analysed during the current study are publicly available at <https://github.com/ana-vranic/Stack-Exchange-communities>.

Declarations

Competing interests

The authors declare that they have no competing interests.

Author contribution

AV, AT, AA, MMD designed the research. AV, AT and AA collected the data and performed data analysis. All authors wrote and edited the final manuscript. All authors read and approved the final manuscript.

Author details

¹Institute of Physics Belgrade, University of Belgrade, Pregrevice 118, Belgrade, Serbia. ²Department of Sociology, Faculty of Philosophy, University of Novi Sad, Novi Sad, Serbia. ³Two desperados, Belgrade, Serbia.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 July 2022 Accepted: 14 February 2023 Published online: 24 February 2023

References

- Leydesdorff L (2001) In: A sociological theory of communication: the self-organization of the knowledge-based society. Universal-Publishers, USA. <https://doi.org/10.1108/jd.2002.58.1.106.2>
- Leydesdorff L (2012) The triple helix, quadruple helix, ..., and an n-tuple of helices: explanatory models for analyzing the knowledge-based economy? *J Knowl Econ* 3(1):25–35. <https://doi.org/10.1007/s13132-011-0049-4>
- Lipkova H, Landová H, Jarolímková A (2017) Information literacy vis-a-vis epidemic of distrust. In: European conference on information literacy. Springer, Berlin, pp 833–843
- Lucassen T, Schraagen JM (2012) Propensity to trust and the influence of source and medium cues in credibility evaluation. *J Inf Sci* 38(6):566–577
- Abraham B, Parigi P, Gupta A, Cook KS (2017) Reputation offsets trust judgments based on social biases among airbnb users. *Proc Natl Acad Sci* 114(37):9848–9853
- Dankulov MM, Melnik R, Tadić B (2015) The dynamics of meaningful social interactions and the emergence of collective knowledge. *Sci Rep* 5(1):1–10. <https://doi.org/10.1038/srep12197>
- Saxena A, Reddy H (2021) Users roles identification on online crowdsourced q&a platforms and encyclopedias: a survey. *J Comput Soc Sci* 1–33. <https://doi.org/10.1007/s42001-021-00125-9>
- Santos T, Walk S, Kern R, Strohmaier M, Helic D (2019) Activity archetypes in question-and-answer (q&a) websites—a study of 50 stack exchange instances. *ACM Trans Soc Comput* 2(1):1–23. <https://doi.org/10.1145/3301612>
- Slag R, de Waard M, Bacchelli A (2015) 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. IEEE, pp 458–461. <https://doi.org/10.1109/MSR.2015.63>
- Chhabra A, Iyengar SRS (2020) Activity-selection behavior of users in stackexchange websites. In: Companion proceedings of the web conference 2020, pp 105–106. <https://doi.org/10.1145/3366424.3382720>
- Dev H, Geigle C, Hu Q, Zheng J, Sundaram H (2018) The size conundrum: why online knowledge markets can fail at scale. In: Proceedings of the 2018 world wide web conference, pp 65–75. <https://doi.org/10.1145/3178876.3186037>
- Santos T, Walk S, Kern R, Strohmaier M, Helic D (2019) Self-and cross-excitation in stack exchange question & answer communities. In: The world wide web conference, pp 1634–1645. <https://doi.org/10.1145/3308558.3313440>
- Oliver PE, Marwell G (2001) Whatever happened to critical mass theory? A retrospective and assessment. *Social Theory* 19(3):292–311. <https://doi.org/10.1111/0735-2751.00142>
- Smiljanić J, Mitrović Dankulov M (2017) Associative nature of event participation dynamics: a network theory approach. *PLoS ONE* 12(2):0171565. <https://doi.org/10.1371/journal.pone.0171565>
- Török J, Kertész J (2017) Cascading collapse of online social networks. *Sci Rep* 7(1):16743. <https://doi.org/10.1038/s41598-017-17135-1>
- Lórinz L, Koltai J, Győr AF, Takács K (2019) Collapse of an online social network: burning social capital to create it? *Soc Netw* 57:43–53. <https://doi.org/10.1016/j.socnet.2018.11.004>
- Wasko MM, Faraj S (2005) Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Q* 29(1):35–57. <https://doi.org/10.2307/25148667>
- Hung S-Y, Durcikova A, Lai H-M, Lin W-M (2011) The influence of intrinsic and extrinsic motivation on individuals' knowledge sharing behavior. *Int J Hum-Comput Stud* 69(6):415–427. <https://doi.org/10.1016/j.ijhcs.2011.02.004>
- Rode H (2016) To share or not to share: the effects of extrinsic and intrinsic motivations on knowledge-sharing in enterprise social media platforms. *J Inf Technol* 31(2):152–165. <https://doi.org/10.1057/jit.2016.8>

20. Kairam SR, Wang DJ, Leskovec J (2012) 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, pp 673–682. <https://doi.org/10.1145/2124295.2124374>
21. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) Complex networks: structure and dynamics. *Phys Rep* 424(4–5):175–308. <https://doi.org/10.1016/j.physrep.2005.10.009>
22. Gallagher RJ, Young J-G, Welles BF (2021) A clarified typology of core-periphery structure in networks. *Sci Adv* 7(12):9800. <https://doi.org/10.1126/sciadv.abc9800>
23. Melnikov A, Lee J, Rivera V, Mazzara M, Longo L (2018) Towards dynamic interaction-based reputation models. In: 2018 IEEE 32nd international conference on Advanced Information Networking and Applications (AINA), pp 422–428. <https://doi.org/10.1109/AINA.2018.00070>
24. Wei X, Chen W, Zhu K (2015) Motivating user contributions in online knowledge communities: virtual rewards and reputation. In: 2015 48th Hawaii international conference on system sciences. IEEE, pp 3760–3769. <https://doi.org/10.1109/HICSS.2015.452>
25. Yanovsky S, Hoernle N, Lev O, Gal K (2019) One size does not fit all: badge behavior in q&a sites. In: Proceedings of the 27th ACM conference on user modeling, adaptation and personalization, pp 113–120. <https://doi.org/10.1145/3320435.3320438>
26. Santos T, Burghardt K, Lerman K, Helic D (2020) Can badges Foster a more welcoming culture on q&a boards? In: Proceedings of the international AAAI conference on web and social media, vol 14, pp 969–973
27. Bornfeld B, Rafaeli S (2019) When interaction is valuable: feedback, churn and survival on community question and answer sites: the case of stack exchange. In: Proceedings of the 52nd Hawaii international conference on system sciences
28. Kang M (2021) Motivational affordances and survival of new askers on social q&a sites: the case of stack exchange network. *Journal of the Association for Information Science and Technology*. <https://doi.org/10.1002/asi.24548>
29. Ahmed S, Yang S, Johri A (2015) Does online q&a activity vary based on topic: a comparison of technical and non-technical stack exchange forums. In: Proceedings of the second (2015) ACM conference on learning@ scale, pp 393–398. <https://doi.org/10.1145/2724660.2728701>
30. Chen G, Mok L (2021) Characterizing growth and decline in online ux communities. In: Extended abstracts of the 2021 CHI conference on human factors in computing systems, pp 1–7. <https://doi.org/10.1145/3411763.3451646>
31. Posnett D, Warburg E, Devanbu P, Filkov V (2012) Mining stack exchange: expertise is evident from initial contributions. In: 2012 international conference on social informatics. IEEE, pp 199–204. <https://doi.org/10.1109/SocialInformatics.2012.67>
32. Pal A, Chang S, Konstan JA (2012) Evolution of experts in question answering communities. In: Sixth international AAAI conference on weblogs and social media
33. Oliveira N, Muller M, Andrade N, Reinecke K (2018) The exchange in stackexchange: Divergences between stack overflow and its culturally diverse participants. *Proc ACM Hum-Comput Interact* 2(CSCW):1–22. <https://doi.org/10.1145/3274399>
34. Dover Y, Kelman G (2018) Emergence of online communities: empirical evidence and theory. *PLoS ONE* 13(11):0205167. <https://doi.org/10.1371/journal.pone.0205167>
35. Han X, Cao S, Shen Z, Zhang B, Wang W-X, Cressman R, Stanley HE (2017) Emergence of communities and diversity in social networks. *Proc Natl Acad Sci* 114(11):2887–2891. <https://doi.org/10.1073/pnas.1608164114>
36. Kleineberg K-K, Boguñá M (2015) Digital ecology: coexistence and domination among interacting networks. *Sci Rep* 5(1):1–11. <https://doi.org/10.1038/srep10268>
37. Oliver P, Marwell G, Teixeira R (1985) A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *Am J Sociol* 91(3):522–556. <https://doi.org/10.1086/228313>
38. Dunning D, Anderson JE, Schlösser T, Ehlebracht D, Fetchenhauer D (2014) Trust at zero acquaintance: more a matter of respect than expectation of reward, vol 107 pp 122–141. <https://doi.org/10.1037/a0036673>
39. Dunning D, Fetchenhauer D, Schlösser T (2019) Why people trust: solved puzzles and open mysteries. *Curr Dir Psychol Sci* 28(4):366–371. <https://doi.org/10.1177/0963721419838255>
40. Schilke O, Reimann M, Cook KS (2021) Trust in Social Relations. *Annu Rev Sociol* 47(1):239–259. <https://doi.org/10.1146/annurev-soc-082120-082850>
41. McEvily B, Zaheer A, Soda G (2021) Network trust. In: Gillespie N, Fulmer A, Lewicki R (eds) *Understanding trust in organizations*. Taylor & Francis. <https://doi.org/10.4324/9780429449185>
42. Aberer K, Despotovic Z (2001) Managing trust in a peer-2-peer information system. In: CIKM'01. Association for Computing Machinery, New York, pp 310–317. <https://doi.org/10.1145/502585.502638>
43. Duma C, Shahmehri N, Caronni G (2005) Dynamic trust metrics for peer-to-peer systems. In: 16th international workshop on database and expert systems applications (DEXA'05). IEEE, pp 776–781. <https://doi.org/10.1109/DEXA.2005.80>
44. Tschannen-Moran M, Hoy W (2000) A multidisciplinary analysis of the nature, meaning, and measurement of trust. In: *Review of educational research*, vol 70. American Educational Research Association, pp 547–593. <https://doi.org/10.3102/00346543070004547>
45. Dover Y, Goldenberg J, Shapira D (2020) Sustainable online communities exhibit distinct hierarchical structures across scales of size. *Proc R Soc A* 476(2239):20190730. <https://doi.org/10.1098/rspa.2019.0730>
46. Orsini C, Dankulov MM, Colomer-de-Simón P, Jamakovic A, Mahadevan P, Vahdat A, Bassler KE, Toroczkai Z, Boguñá M, Caldarelli G et al (2015) Quantifying randomness in real networks. *Nat Commun* 6(1):8627. <https://doi.org/10.1038/ncomms9627>
47. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) 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, pp 44–54. <https://doi.org/10.1145/1150402.1150412>
48. Centola D, Eguíluz VM, Macy MW (2007) Cascade dynamics of complex propagation. *Phys A, Stat Mech Appl* 374(1):449–456. <https://doi.org/10.1016/j.physa.2006.06.018>
49. Bollobás B, Riordan OM (2003) Mathematical results on scale-free random graphs. In: *Handbook of graphs and networks: from the genome to the Internet*, pp 1–34

50. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174. <https://doi.org/10.1016/j.physrep.2009.11.002>
51. Saramäki J, Moro E (2015) From seconds to months: an overview of multi-scale dynamics of mobile telephone calls. *Eur Phys J B* 88(6):1–10. <https://doi.org/10.1140/epjb/e2015-60106-6>
52. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Sci* 1(1):1. <https://doi.org/10.1140/epjds4>
53. Barrat A, Gelardi V, Le Bail D, Claidiere N (2021) From temporal network data to the dynamics of social relationships. *Proc R Soc Lond B, Biol Sci* 288:20211164. <https://doi.org/10.1098/rspb.2021.1164>
54. Arnold NA, Steer B, Hafnaoui I, Parada GHA, Mondragon RJ, Cuadrado F, Clegg RG (2021) Moving with the times: investigating the alt-right network gab with temporal interaction graphs. *Proc ACM Hum-Comput Interact* 5(CSCW2) 447. <https://doi.org/10.1145/3479591>
55. Yashkina E, Pinigin A, Lee J, Mazzara M, Adekotujo AS, Zubair A, Longo L (2019) Expressing trust with temporal frequency of user interaction in online communities. In: *Advanced information networking and applications*. Springer, Cham. https://doi.org/10.1007/978-3-030-15032-7_95

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

PAPER: Interdisciplinary statistical mechanics

Universal growth of social groups: empirical analysis and modeling

Ana Vranic^{1,*}, Jelena Smiljanic^{1,2} and Marija Mitrović Dankulov¹

¹ Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia

² Integrated Science Lab, Department of Physics, Umeå University, SE-901 87 Umeå, Sweden

E-mail: ana.vranic@ipb.ac.rs, jelena.smiljanic@ipb.ac.rs and marija.mitrovic.dankulov@ipb.ac.rs

Received 15 June 2022

Accepted for publication 28 October 2022

Published 7 December 2022



Online at stacks.iop.org/JSTAT/2022/123402
<https://doi.org/10.1088/1742-5468/aca0e9>

Abstract. Social groups are fundamental elements of any social system. Their emergence and evolution are closely related to the structure and dynamics of a social system. Research on social groups was primarily focused on the growth and the structure of the interaction networks of social system members and how members' group affiliation influences the evolution of these networks. The distribution of groups' size and how members join groups has not been investigated in detail. Here we combine statistical physics and complex network theory tools to analyze the distribution of group sizes in three data sets, Meetup groups based in London and New York and Reddit. We show that all three distributions exhibit log-normal behavior that indicates universal growth patterns in these systems. We propose a theoretical model that combines social and random diffusion of members between groups to simulate the roles of social interactions and members' interest in the growth of social groups. The simulation results show that our model reproduces growth patterns observed in empirical data. Moreover, our analysis shows that social interactions are more critical for the diffusion of members in online groups, such as Reddit, than in offline groups, such as Meetup. This work shows that social groups follow universal growth mechanisms that need to be considered in modeling the evolution of social systems.

*Author to whom any correspondence should be addressed.

Keywords: network dynamics, random graphs, networks, scaling in socio-economic systems, stochastic processes

Contents

1. Introduction2

2. Data4

3. Empirical analysis of social group growth5

4. Model8

5. Results11

 5.1. Model properties11

 5.2. Modeling real systems12

6. Discussion and conclusions16

Acknowledgments18

References18

1. Introduction

The need to develop methods and tools for their analysis and modeling comes with massive data sets. Methods and paradigms from statistical physics have proven to be very useful in studying the structure and dynamics of social systems [1]. The main argument for using statistical physics to study social systems is that they consist of many interacting elements. Due to this, they exhibit different patterns in their structure and dynamics, commonly known as *collective behavior*. While various properties can characterize a social system’s building units, only a few 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 [2], or scientific citations [3]. It indicates the existence of the universal mechanisms that govern the dynamics of the system [1].

Social groups, informal or formal, are mesoscopic building elements of every socio-economic system that direct its emergence, evolution, and disappearance [4]. The examples span from countries, economies, and science to society. Settlements, villages, towns, and cities are formal and highly structured social groups of countries. Their organization and growth determine the functioning and sustainability of every society [5]. Companies are the building blocks of an economic system, and their dynamics are essential indicators of the level of its development [6]. 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 [7]. The membership of

J. Stat. Mech. (2022) 123402

individuals in various social groups, online and offline, can be essential when it comes to the quality of their life [8–10]. Therefore, it is not surprising that the social group emergence and evolution are at the center of the attention of many researchers [11–14].

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 [15] and remain its active members [7, 16]. The study on LiveJournal [15] 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 [7]. The sense of belonging of an individual in social groups is achieved through two main mechanisms [16]: expanding the social circle at the beginning of joining the group and strengthening the existing connections in the later phase. Analysis of the evolution of large-scale social networks has shown that edge locality plays a critical role in the growth of social networks [17]. The dynamics of social groups depend on their size [18]. Small groups are more cohesive with continued long-term, while large groups change their active members constantly [18]. 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 group growth. However, how the growth mechanisms influence the distribution of members of one social system among groups is yet to be understood.

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 [19]. However, the distribution of company sizes follows log-normal behavior and remains stable over decades [20, 21]. Analysis of the cities' sizes shows that all cities' distribution also follows a log-normal distribution [22]. In contrast, the distribution of the largest cities resembles Zipf's distribution [23].

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 produce companies' and cities' observed log-normal size distribution. However, the analysis of the growth rate of the companies [20] has shown that growth mechanisms are different from those assumed by Gibrat. In addition, the analysis of the growth of the online social networks showed that the population size and spatial factors do not determine population growth, and it deviates from Gibrat's law [24]. Other mechanisms, for instance, growth through diffusion, have been used to model and predict rapid group growth [25]. 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 three data sets: Meetup groups based in London and New York 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 [15, 19]. We show that the model can reproduce size and growth rate distributions for both studied systems. Moreover, the model is flexible and can produce a broad set of log-normal size distributions depending on the value of model parameters.

The paper is organized as follows: in section 2 we describe the data, while in section 3 we present our empirical results. In section 4 we introduce model parameter and principles. In section 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 6, we present concluding remarks and discuss our results.

2. Data

We analyze the growth of social groups from two widely used online platforms: Reddit and Meetup. Reddit³ enables sharing of diverse web content, and members of this platform interact exclusively online through posts and comments. The Meetup⁴ 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 data science professionals. Due to their specific activity patterns—events where members meet face-to-face—Meetup groups are geographically localized, and interactions between members are primarily offline.

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 selected subreddits created between 2006 and 2011 that were active in 2017 and followed their growth from their beginning until 2011. The considered dataset contains 17073 subreddits with 2195 677 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 the timestamp when a member made a post for the first time. Finally, in the dataset, we include only subreddits active for at least two months.

The Meetup data were downloaded in 2018 using public API. The Meetup platform was launched in 2003, and when 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 in a period between 2003 and 2017 in big cities, London and New York, where the Meetup platform achieved considerable popularity. We considered groups active for at least two months. There were 4673 groups with 831 685 members in London and 4752 groups with 1059 632 members in New York. In addition, we extracted the ids of group

³<https://reddit.com/>.

⁴www.meetup.com.

members, the information about organized events, and which members attended these events. Based on this, we obtained the date when a member joined a group, the first time she participated in a group event.

For all systems, we extracted the timestamp when the member joined the group. Each data set has a form (u_{id}, g_{id}, t_i) , representing the connection between users and groups. When the system has two separate partitions, the natural extension is a bipartite network where links are drawn between nodes of different sets, indicating the user's memberships. The degree of group nodes is exactly the group size. Having the temporal component in data, we can follow the evolution of the network. 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 each group. The time step for all three data sets 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. $S_i(t) = \sum_{k=t_0}^{k=t} N_i(k)$, where t_0 is the time step in which the group i was created. Once the member joins the group, it has an active status by default, which remains permanent. 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 $R_i(t) = \log(S_i(t)/S_i(t-1))$.

While the forms of communication between members and activities that members engage in differ for considered systems, some common properties exist between them. Members can form new groups and join the existing ones. 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 on Reddit and Meetup.

3. Empirical analysis of social group growth

Figure 1 summarizes the properties of the groups in Meetup and Reddit systems. The number of groups grows exponentially over time. Nevertheless, we notice that Reddit has a substantially larger number of groups than Meetup. The Reddit groups are prone to engage more members in a shorter period. The size of the Meetup groups ranges 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 million. The distributions of normalized group sizes follow the log-normal distribution (see table S1 and figure S1 in SI)

$$P(S) = \frac{1}{\frac{S}{S_0} \sigma \sqrt{2\pi}} \exp\left(-\frac{\left(\ln\left(\frac{S}{S_0}\right) - \mu\right)^2}{2\sigma^2}\right), \quad (1)$$

where S is the group size, S_0 is the average group size in the system, and μ and σ are parameters of the distribution. We used *power-law* package [26] to fit equation (1) to empirical 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 $\sigma = 3.07$.

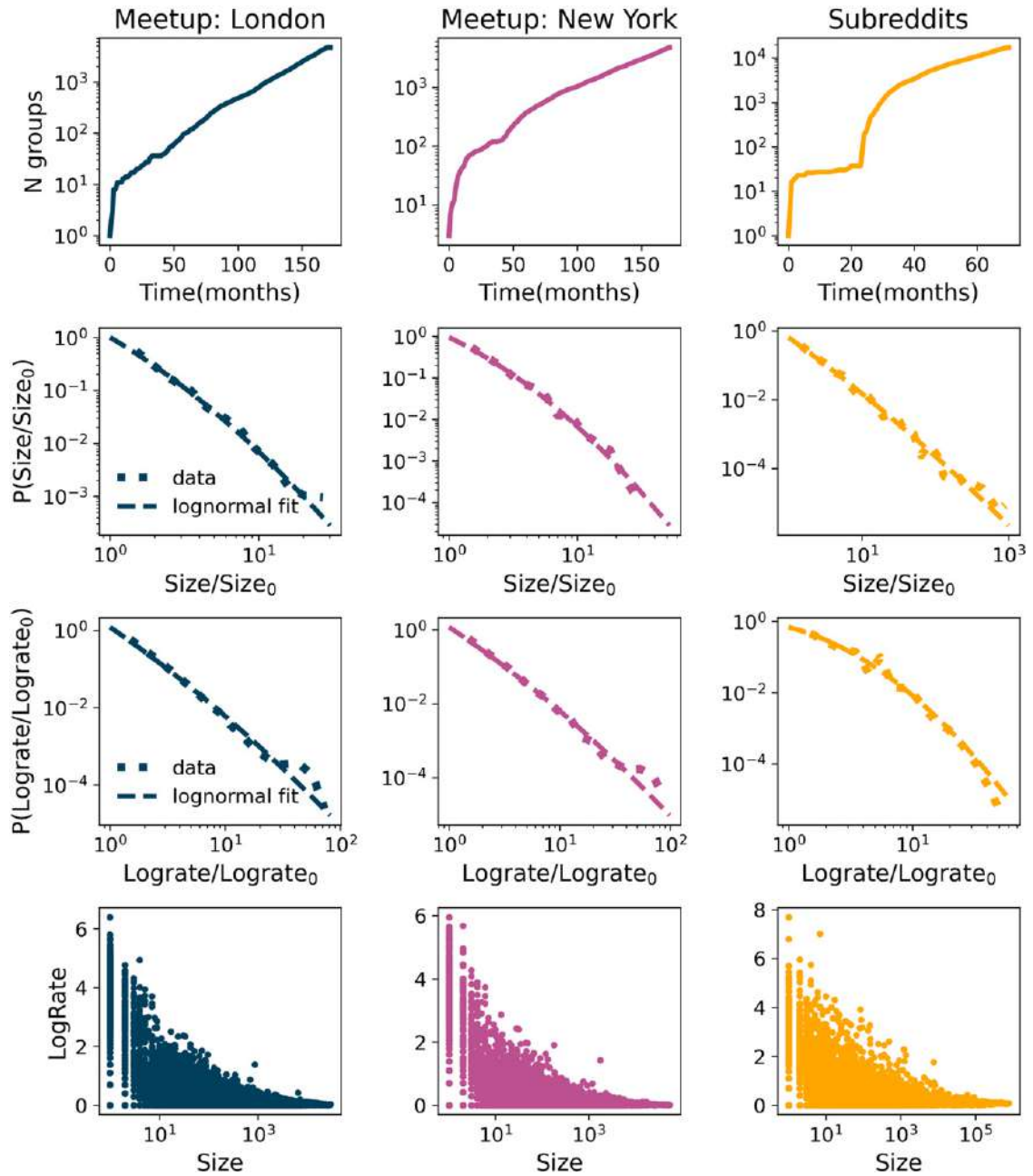


Figure 1. The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London and New York and subreddits. The number of groups grows exponentially over time, while the group size distributions, and log-rates distributions follow log-normal. Logrates depend on the size of the group, implying that the growth cannot be explained by Gibrat law.

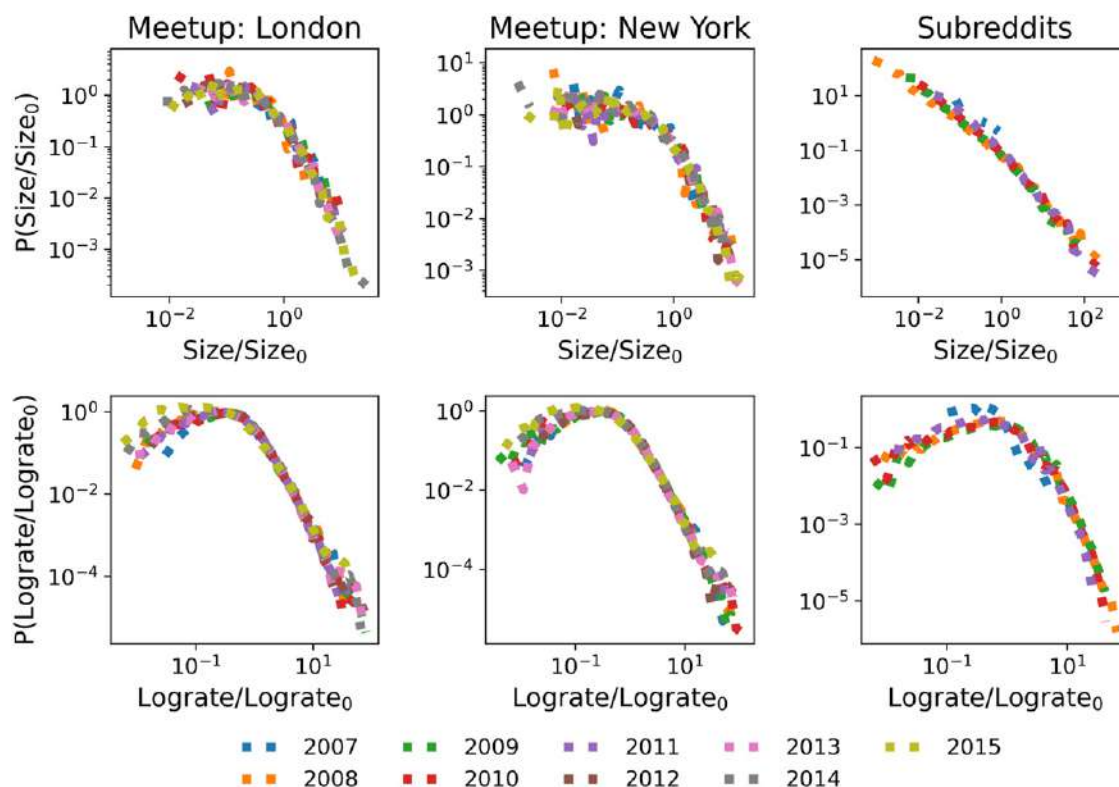


Figure 2. The figure shows the groups' sizes distributions and log-rates distributions. Figures in the top panels show the distribution of normalized sizes of groups created in the same year. Distributions for the same system and different years follow same log-normal distribution indicating existence of universal growth patterns.

Multiplicative processes can generate the log-normal distributions [27]. 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 number. The Gibrat law states that growth rates r are uncorrelated and do not depend on the current size. To describe the growth of social groups, we calculate the logarithmic growth rates $R_i(t)$. According to Gibrat law the distribution of logarithmic growth rates is normal, or, as it is shown in many studies, it is better explained with Laplacian ('tent-shaped') distribution [28, 29]. In figure 1 we show the distributions of log-rates for all three data sets. Log-rates are very well approximated with a log-normal distribution. Furthermore, the bottom panels of figure 1 show that log-rates are not independent of group size. Figure 1 shows that these findings imply that the growth of Meetup and Reddit groups violates the basic assumptions of Gibrat's law [30, 31] and that it cannot be explained as a simple multiplicative process.

We are considering a relatively significant period for online groups. The fast expansion of information communications technologies (ICT) changed how members access online systems. With the use of smartphones, online systems became more available,

which led to the exponential growth of ICTs systems and potential change in the mechanisms that influence the social groups' growth. For these reasons, we aggregate groups according to the year they were founded for each of the three data sets and look at the distributions of their sizes at the end of 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 created in a year $y \langle S^y \rangle$. We normalize the size of the groups originating in year y with the corresponding average size $s_i^y = S_i^y / \langle S^y \rangle$ and calculate the distribution of the normalized sizes for each year. The distribution of normalized sizes for all years and data sets is shown in figure 2. All distributions exhibit log-normal behavior. Furthermore, the distributions for the same data set and different years follow a universal curve with the same value of parameters μ and σ . The universal behavior is observed for the distribution of normalized log-rates as well, see figure 2 (bottom panels). These results indicate that the growth of the social groups did not change due to the increased growth of members in systems. Furthermore, it implies that the growth is independent of the size of the whole data set.

4. Model

The growth of social groups cannot be explained by 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 [19, 25]. Individuals' interests and the need to discover new content or activity also influence the 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 [32, 33] and the structure of social interactions [34]. The number of social groups in the social systems is not constant. They are constantly created and destroyed.

In [19], the authors propose the co-evolution model of the growth of social networks. In this model, the authors assume that the social system evolves through the co-evolution of two networks: a network of social contacts between members and a network of members' affiliations with groups. This model addresses the problem of the 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 for larger groups. If a 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.

In [19], the authors demonstrate that mechanisms postulated in the model could reproduce the power-law distribution of group sizes observed for some social networks. However, as illustrated in section 3, the distribution of group sizes in real systems is not necessarily power-law. Our rigorous empirical analysis shows that the distribution of social group sizes exhibits log-normal behavior. To fill the gap in understanding how social groups in the social system grow, we propose a model of group growth that

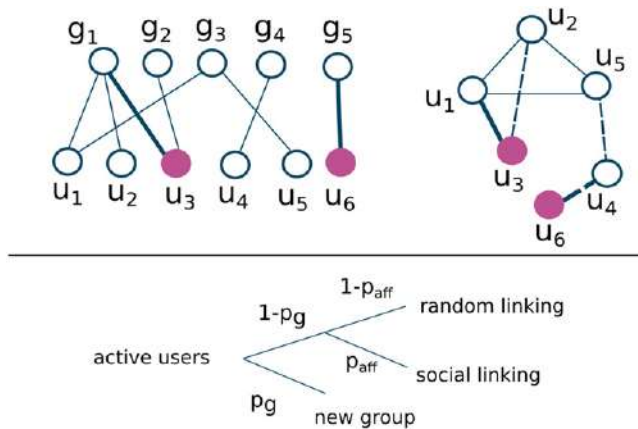


Figure 3. The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model 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 .

combines random and social diffusion between groups but follows different rules than the co-evolution model [19].

Figure 3 shows a schematic representation of our model. Similar to the co-evolution model [19], we represent a social system with two evolving networks, see figure 3. One network is a bipartite network that 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 a 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: the arrival of new members, the creation of new groups, and members joining groups. In bipartite networks, links only exist between 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. However, previous research on this subject [15, 16, 19] 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 by adding new members to the set V_U and creating 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 the bipartite and social network of members with adjacency matrices B and A . The element of the matrix B_{ug} equals one if member u

is affiliated with group g , and zero otherwise. In matrix A , the element $A_{u_1 u_2}$ equals one if members u_1 and u_2 are connected and zero otherwise. The neighborhood \mathcal{N}_u of member u is a set of groups with which the member is affiliated. On the other hand, the neighborhood \mathcal{N}_g of a group g is a set of members affiliated with that group. The size S_g of set \mathcal{N}_g equals to the size of the group 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 [25], i.e. to select a group according to her social contacts. Not all members from setting V_U are active in each time step. Only a subset of existing members is active in each time step. The activity of old members is a stochastic process determined by parameter p_a ; every old member is activated with probability p_a . Old members are 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 creating 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 choose a 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 the group g equals the overlap of members affiliated with a group g and social contacts of member u , and is calculated according to

$$s_{ug} = \sum_{u_1 \in \mathcal{N}_g} A_{uu_1}. \quad (2)$$

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 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 the bipartite network between a member u and a group g . At the same time, the 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 make X new links in-network \mathcal{G} between member u and X members from a 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 a higher number of active members and thus faster growth of the number of links in both networks and 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 central nodes and groups as leaves. On the other hand, $p_g = 0$ means that there is no creation of new groups, and the bipartite network only grows through adding new members and creating new links between members and groups.

Parameter p_{aff} determines the importance of social diffusion. $p_{\text{aff}} = 0$ means that social connections are irrelevant, and the group choice is random. On the other hand, $p_{\text{aff}} = 1$ means that only social contacts become important for group selection.

Several differences exist between the model presented in this work and the co-evolution model [19]. In our model, p_{aff} is constant and the same for all members. In the co-evolution model, this probability depends on members' degrees. The members are activated in our model with probability p_a . In contrast, in the 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 an exponential distribution. In the co-evolution model, the number of social contacts members have within the group is irrelevant to its selection. On the other hand, in our model, members tend to choose groups more often in which there is a greater number of social contacts. While in our model, in the case of a random selection of a group, a member selects with equal probability a group that she is not affiliated with, in the co-evolution model, the choice of group is preferential.

5. Results

The distribution of group sizes produced by our and co-evolution models significantly differ. The distribution of group sizes in the co-evolution model is a power-law. Our model enables us to create groups with log-normal size distribution and expand classes of social systems that can be modeled.

5.1. Model properties

First, we explore the properties of size distribution depending on parameters p_g and p_{aff} , for the fixed value of activity parameter p_a and constant number of members added in each step $N(t) = 30$. When the group is created, its size $S(t_0) = 1$, so the group creator cannot make new social connections until new members arrive. While a group has less than X members, new users will make social connections with all available members in the group. After the group size reaches the threshold of X members, a new user creates X connections. Our detailed analysis of the results for different parameter values X shows that these results are independent of their value. We set the value of parameter X to 25 for all simulations presented in this work. Our detailed analysis of the results for different parameter values X shows that these results are independent of their value.

Figure 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 figure 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 in figure 4, the size distribution begins to deviate from log-normal distribution. The higher the value of parameter p_g , the total number of groups grows faster. For $p_g = 0.5$, half of the active members in each time step create a group, and the number of groups increases fast. How members are distributed in these groups

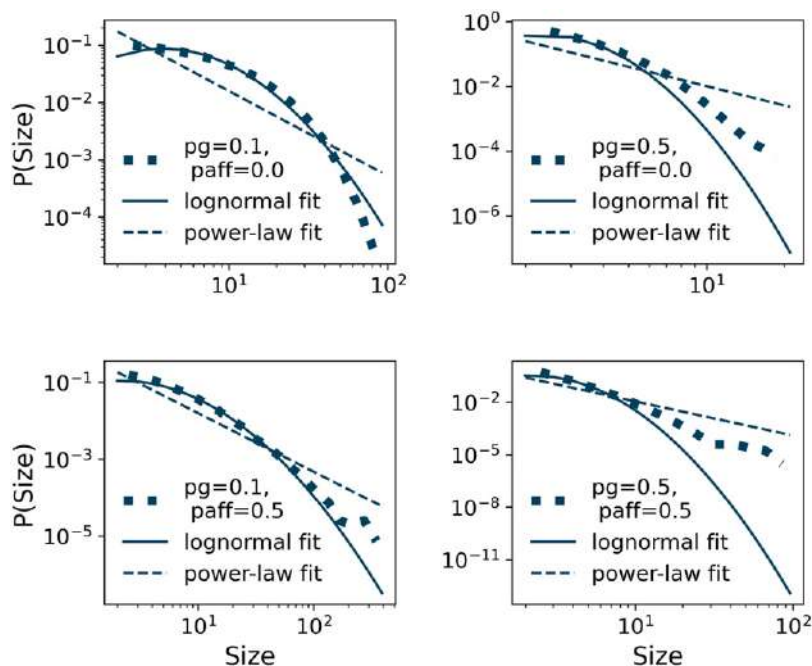


Figure 4. The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

depends on the parameter p_{aff} value. When $p_{aff} = 0$, social connections are irrelevant to the group's choice, and members select groups randomly. The obtained distribution slightly deviates from log-normal, especially for large group sizes. In this case, large group sizes become more probable than in the case of the log-normal distribution. The non-zero value of parameter p_{aff} means that the choice of a group becomes dependent on social connections. When a member chooses a group according to her social connections, larger groups have a higher probability of being affiliated with the 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.

The top panel of figure S3 in SI shows how the shape of distribution is changing with the value of parameter p_{aff} and fixed values of $p_a = 0.1$ and $p_g = 0.1$. Preferential selection groups according to their size instead of one where a member selects a group with equal probability leads to a drastic change in the shape of the distribution, bottom panel figure S3 in SI. As is to be expected, the distribution of group sizes with preferential attachment follows power-law behavior.

5.2. Modeling real systems

The social systems do not grow at a constant rate. In [34], the 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

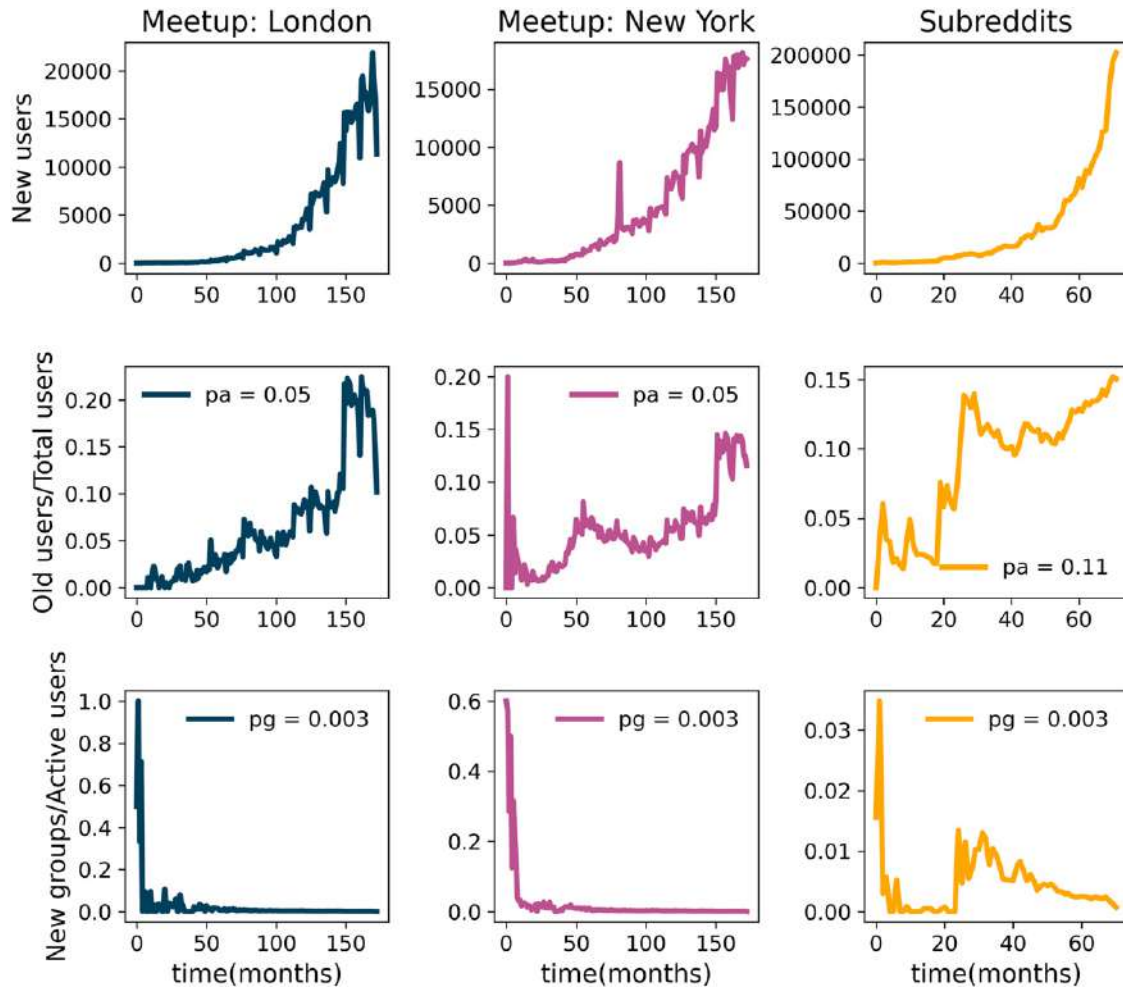


Figure 5. The time series of the number of new members (top panels). The time series of the ratio between several old active members and total members in the system (middle panels); its median value approximates the parameter p_a , the probability that the user is active. The bottom panels show the time series of the ratio between new groups and active members; its median value approximates the probability that active users create a new group, parameter p_g .

and Reddit to simulate the growth of the social groups in these systems. Figure 5 (top) shows the time series of the number of new members that join each of the considered systems each month. All three data sets have relatively low growth at the beginning, and then the growth accelerates as the system becomes more popular.

We also use empirical data to estimate p_a , p_g and p_{aff} . The data can approximate the probability that old members are active p_a and that new groups are created p_g . Activity parameter p_a is the ratio between the number of old members active in month t and the total number of members in the system at time t . Figure 5 (middle) shows the variation of parameter p_a during the considered time interval for each system. The value of this parameter fluctuates between 0 and 0.2 for London and New York based Meetup

Table 1. Jensen Shannon divergence between group sizes distributions from model and data. In the model we vary affiliation parameter p_{aff} and find its optimal value (bold text).

p_{aff}	JS cityLondon	JS cityNY	JS reddit2012
0.1	0.0161	0.0097	0.002 41
0.2	0.0101	0.0053	0.002 05
0.3	0.0055	0.0026	0.001 59
0.4	0.0027	0.0013	0.001 04
0.5	0.0016	0.0015	0.000 74
0.6	0.0031	0.0035	0.000 48
0.7	0.0085	0.0081	0.000 39
0.8	0.0214	0.0167	0.000 34
0.9	0.0499	0.0331	0.000 47

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 and 80 months for Reddit systems. 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 5 bottom row shows the evolution of parameter p_g for the considered systems. The p_g in month t is estimated as the ratio between the groups created in month $tNg_{\text{new}}(t)$ and the total number of groups in that month $Ng_{\text{new}}(t) + Ng_{\text{old}}(t)$, i.e. $p_g(t) = \frac{Ng_{\text{new}}(t)}{Ng_{\text{new}}(t) + Ng_{\text{old}}(t)}$. We see from figure 5 that $p_g(t)$ has relatively high values at the beginning of the system's existence. This is not surprising. Initially, these systems have a relatively small number of groups and often cannot meet the needs of the content of all their members. As the time passes, the number of groups and content scope within the system grows, and members no longer have a high need to create new groups. Figure 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 the median value during 170 months for Meetup and 80 months for Reddit. For all three systems p_g has the value of 0.003.

The affiliation parameter p_{aff} cannot estimate directly from the empirical data. For these reasons, we simulate the growth of social groups for each data set 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} . We compare the distribution of group sizes obtained from simulations for different values of p_{aff} with ones obtained from empirical analysis using Jensen Shannon (JS) divergence. The JS divergence [35] between two distributions P and Q is defined as

$$JS(P, Q) = H\left(\frac{P + Q}{2}\right) - \frac{1}{2}(H(P) + H(Q)) \quad (3)$$

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. Table 1

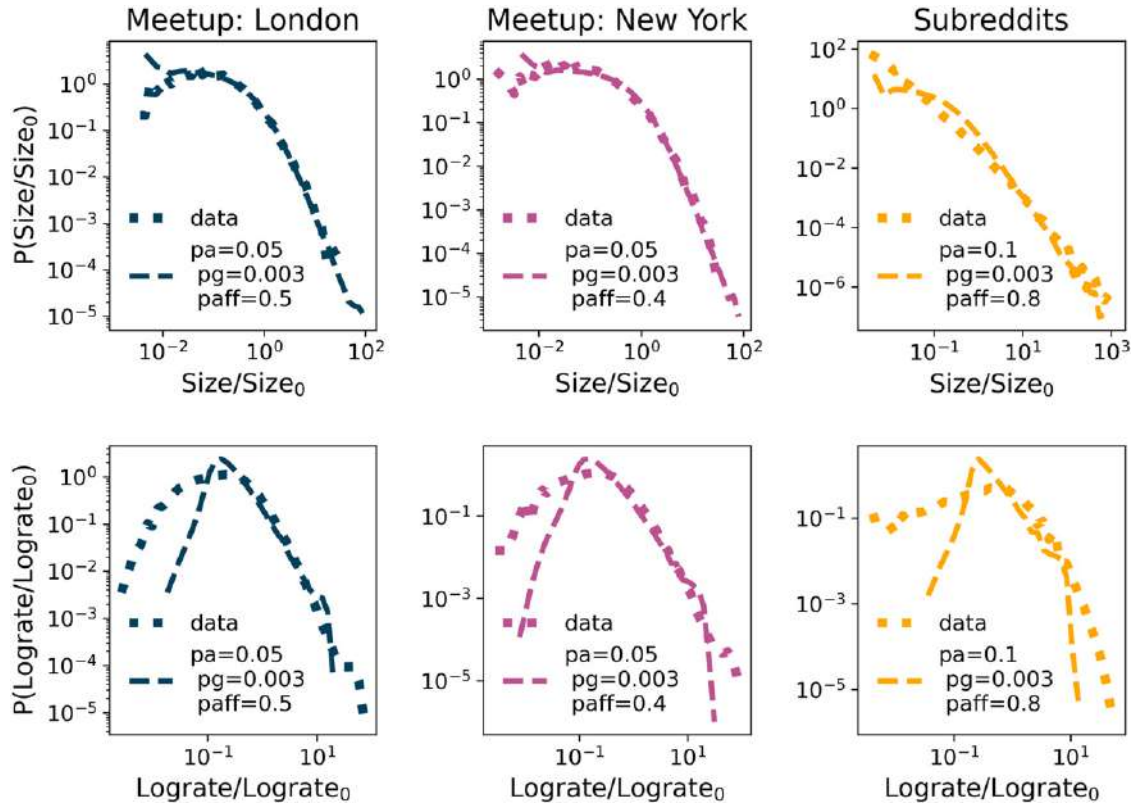


Figure 6. The comparison between empirical and simulation distribution for group sizes (top panels) and log-rates (bottom panels).

shows the value of JS divergence for all three data sets. We see that for London based Meetup groups the affiliation parameter is $p_{\text{aff}} = 0.5$, for New York groups $p_{\text{aff}} = 0.4$, while the affiliation parameter for Reddit $p_{\text{aff}} = 0.8$. Our results show that social diffusion is important in all three data sets. 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 6 compares the empirical and simulation distribution of group sizes for considered systems. We see that empirical distributions for Meetup groups based in London and New York are well reproduced by the model and chosen values of parameters. In the case of Reddit, the distribution is broad, and the model reproduces the tail of the distribution well. Figure S2 and table S2 in SI confirm that the distribution of group sizes follow a log-normal distribution.

The bottom row of figure 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 data sets are well emulated by the ones obtained from the model. The deviations we observe are the most likely consequence of using median values of parameters p_a , p_g , and p_{aff} .

6. Discussion and conclusions

The results of empirical analysis show that there are universal growth rules that govern the growth of social systems. We analysed the growth of social groups for three data sets, Meetup groups located in London and New York and Reddit. We showed that the distribution of group sizes has log-normal behaviour. The empirical distributions of normalised sizes of groups created in different years in a single system fall on top of each other, following the same log-normal distributions. Due to a limited data availability, we only study three data sets which may affect the generality of our results. However, the substantial differences between Reddit and Meetup social systems when it comes to their popularity, size and purpose, demonstrate that observed growth patterns are universal.

Even though the log-normal distribution of group sizes can originate from the proportional growth model, Gibrat law, we show that it does not apply to the growth of online social groups. The monthly growth rates are log-normally distributed and dependent on the size of a group. Gibrat law was proposed to describe the growth of various socio-economical systems, including the cities and firms. Recent studies showed that the growth of cities and firms [21, 36, 37] goes beyond Gibrat law. Still, our findings confirm the existence of universal growth patterns, indicating the presence of the general law in the social system's growth.

While the growth of the social groups does not follow the Gibrat law, one could ask whether there are other simple models of social group growth. The basic growth model underlying any log-normal distribution is a multiplicative process. The size of the system in time t is equal to its size in time $t - 1$ multiplied by some factor. In our case, where the groups only grow and do not shrink, the factor has to be larger than one. When we model the growth of real social groups, we need to take into account several factors: (1) social systems grow through the addition of new members; (2) the number of social groups is not constant, it grows with time; (3) one person can be a member of multiple groups at the same time. The simplest model that considers all three factors but disregards social factors, and thus a network structure, would be the one where members randomly choose the groups they will join. The described situation is an extreme case of our model with $p_{\text{aff}} = 0$, see figure 4, top left panel. By setting the values of $p_{\text{aff}} = 0$ and taking the value of $N(t)$ and p_g as an estimate from real data, we can reproduce a log-normal distribution with parameters that do not match empirical data, see table 1. While the distributions of group size in different systems follow log-normal behavior, the parameters of these distributions differ from system to system. This indicates the existence of additional factors in the multiplicative process that govern multiplicative growth. The network effect is crucial in explaining many instances of collective social dynamics, including the person's choice to join a certain group [14]. Here we show that members' diffusion between groups governed by social influence allows us to use the same model to explain the growth of groups in different social systems by tuning its importance.

The model proposed in [19] is able to produce only power-law distributions of group sizes. However, our empirical analysis shows that these distributions can also have a log-normal behavior. Thus, we propose a new model that emulate log-normal distributions. The analysed groups grow through two mechanisms [19]: members join a group that is chosen according to their interests or by social relations with the group's members. The number of members in the system is growing as well as the number of groups. While the processes that govern the growth of social groups are the same, their importance varies among the systems. The distributions for Meetup groups located in the London and New York have similar log-normal distribution parameter values, while for Reddit, the distribution is broader. Numerical simulations further confirm these findings. Different modalities of interactions between their members can explain the observed differences.

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 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.

The purpose of the research presented in this paper was to provide a model of social group growth that can reproduce the log-normal distribution of group sizes in different systems. The model is based on bipartite network dynamics allowing us to study other network properties and compare them to empirical data. The empirical data are limited and only contain explicit information about the connections between groups and their members. The distribution of group sizes is the exact degree distribution of the group partition. We show that these properties are reproduced with our model, see figure 6. When it comes to the degree distribution of members, that is, the number of groups a member is affiliated with, our model does not reproduce this distribution. The number of groups a member is affiliated to is equal to number of her activities. The activity of a member is controlled with probability p_a . In our model, the probability p_a is equal for all members, and thus the emerging degree distribution is exponential [38]. We do not study the properties of the members' partitions in detail, as our focus is on the growth of groups' partitions and mechanisms that influence the members' choice to join the groups. On the other hand, studying how groups are distributed among members could give us insight into what motivates members to be active. Previous work proposed that each member has a lifetime [17], but different linking rules could be considered; for example, p_a could be preferential toward high-degree members, and the age or even social connections of members could be relevant.

The results presented in this paper contribute to our knowledge of the growth of socio-economical systems. The previous study analysed the social systems in which size distributions follow the power-law, which is the consequence of a preferential choice of groups during the random diffusion of members. Our findings show that preferential

selection of groups during social diffusion and uniform selection during random diffusion result in log-normal distribution of groups sizes. Furthermore, we show that broadness of the distribution depends on the involvement of social diffusion in the growth process. Our model increases the number of systems that can be modelled and help us better understand the growth and segmentation of social systems and predict their evolution.

Acknowledgments

We acknowledge funding provided by the Institute of Physics Belgrade, through the grant by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. Numerical simulations were run on the PARADOX-IV supercomputing facility at the Scientific Computing Laboratory, National Center of Excellence for the Study of Complex Systems, Institute of Physics Belgrade.

References

- [1] Castellano C, Fortunato S and Loreto V 2009 Statistical physics of social dynamics *Rev. Mod. Phys.* **81** 591
- [2] Chatterjee A, Mitrović M and Fortunato S 2013 Universality in voting behavior: an empirical analysis *Sci. Rep.* **3** 1–9
- [3] Radicchi F, Fortunato S and Castellano C 2008 Universality of citation distributions: toward an objective measure of scientific impact *Proc. Natl Acad. Sci. USA* **105** 17268–72
- [4] Firth R 2013 *Elements of Social Organisation* (London:Routledge)
- [5] Barthelémy M 2016 *The Structure and Dynamics of Cities* (Cambridge: Cambridge University Press)
- [6] Hidalgo C A and Hausmann R 2009 The building blocks of economic complexity *Proc. Natl Acad. Sci. USA* **106** 10570–5
- [7] Smiljanić J, Chatterjee A, Kauppinen T and Dankulov M M 2016 A theoretical model for the associative nature of conference participation *PLoS One* **11** e0148528
- [8] Montazeri A, Jarvandi S, Haghghat S, Vahdani M, Sajadian A, Ebrahimi M and Haji-Mahmoodi M 2001 Anxiety and depression in breast cancer patients before and after participation in a cancer support group *Patient Educ. Counseling* **45** 195–8
- [9] Davison K P, Pennebaker J W and Dickerson S S 2000 Who talks? The social psychology of illness support groups *Am. Psychol.* **55** 205
- [10] Cho W K T *et al* 2012 The tea party movement and the geography of collective action *Q. J. Pol. Sci.* **7** 105–33
- [11] Aral S and Walker D 2012 Identifying influential and susceptible members of social networks *Science* **337** 337–41
- [12] González-Bailón S, Borge-Holthoefer J and Moreno Y 2013 Broadcasters and hidden influentials in online protest diffusion *Am. Behav. Sci.* **57** 943–65
- [13] Török J, Iniguez G, Yasseri T, San Miguel M, Kaski K and Kertész J 2013 Opinions, conflicts, and consensus: modeling social dynamics in a collaborative environment *Phys. Rev. Lett.* **110** 088701
- [14] Yasseri T, Sumi R, Rung A, Kornai A and Kertész J 2012 Dynamics of conflicts in wikipedia *PLoS One* **7** e38869
- [15] Backstrom L, Huttenlocher D, Kleinberg J and Lan X 2006 Group formation in large social networks: membership, growth, and evolution *Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 44–54
- [16] Smiljanić J and Dankulov M M 2017 Associative nature of event participation dynamics: a network theory approach *PLoS One* **12** e0171565
- [17] Leskovec J, Backstrom L, Kumar R and Tomkins A 2008 Microscopic evolution of social networks *Proc. 14th ACM SIGKDD Int. Conf. Knowledge discovery and data mining* pp 462–70
- [18] Palla G, Barabási A-L and Vicsek T 2007 Quantifying social group evolution *Nature* **446** 664–7
- [19] Zheleva E, Sharara H and Getoor L 2009 Co-evolution of social and affiliation networks *Proc. 15th ACM SIGKDD Int. Conf. Knowledge discovery and data mining* pp 1007–16

- [20] Amaral L A N, Buldyrev S V, Havlin S, Leschhorn H, Maass P, Salinger M A, Stanley H E and Stanley M H R 1997 Scaling behavior in economics: I. Empirical results for company growth *J. Phys. I* **7** 621–33
- [21] Stanley M H R, Amaral L A N, Buldyrev S V, Havlin S, Leschhorn H, Maass P, Salinger M A and Stanley H E 1996 Scaling behaviour in the growth of companies *Nature* **379** 804–6
- [22] González-Val R 2019 Lognormal city size distribution and distance *Econ. Lett.* **181** 7–10
- [23] Fazio G and Modica M 2015 Pareto or log-normal? Best fit and truncation in the distribution of all cities *J. Regional Sci.* **55** 736–56
- [24] Zhu K, Li W, Fu X and Nagler J 2014 How do online social networks grow? *PLoS One* **9** e100023
- [25] Kairam S R, Wang D J and Leskovec J 2012 The life and death of online groups: predicting group growth and longevity *Proc. 5th ACM Int. Conf. Web Search and Data Mining* pp 673–82
- [26] Alstott J, Bullmore E and Plenz D 2014 Powerlaw: a python package for analysis of heavy-tailed distributions *PLoS One* **9** 1–11
- [27] Mitzenmacher M 2004 A brief history of generative models for power law and lognormal distributions *Internet Math.* **1** 226–51
- [28] Mondani H, Holme P and Liljeros F 2014 Fat-tailed fluctuations in the size of organizations: the role of social influence *PLoS One* **9** e100527
- [29] Fu D, Pammolli F, Buldyrev S V, Riccaboni M, Matia K, Yamasaki K and Stanley H E 2005 The growth of business firms: theoretical framework and empirical evidence *Proc. Natl Acad. Sci. USA* **102** 18801–6
- [30] Frasco G F, Sun J, Rozenfeld H D and Ben-Avraham D 2014 Spatially distributed social complex networks *Phys. Rev. X* **4** 011008
- [31] Qian J-H, Chen Q, Han D-D, Ma Y-G and Shen W-Q 2014 Origin of Gibrat law in internet: asymmetric distribution of the correlation *Phys. Rev. E* **89** 062808
- [32] Mitrović M, Paltoglou G and Tadić B 2011 Quantitative analysis of bloggers' collective behavior powered by emotions *J. Stat. Mech.* **P02005**
- [33] Dankulov M M, Melnik R and Tadić B 2015 The dynamics of meaningful social interactions and the emergence of collective knowledge *Sci. Rep.* **5** 1–10
- [34] Vranić A and Dankulov M M 2021 Growth signals determine the topology of evolving networks *J. Stat. Mech.* **2021** 013405
- [35] Briët J and Harremoës P 2009 Properties of classical and quantum Jensen–Shannon divergence *Phys. Rev. A* **79** 052311
- [36] Mansfield E 1962 Entry, Gibrat's law, innovation, and the growth of firms *Am. Econ. Rev.* **52** 1023–51
- [37] Barthelemy M 2019 The statistical physics of cities *Nat. Rev. Phys.* **1** 406–15
- [38] Barabási A-L, Albert R and Jeong H 1999 Mean-field theory for scale-free random networks *Physica A* **272** 173–87

PAPER

Growth signals determine the topology of evolving networks

To cite this article: Ana Vrani and Marija Mitrovi Dankulov *J. Stat. Mech.* (2021) 013405

View the [article online](#) for updates and enhancements.



IOP | ebooks™

Bringing together innovative digital publishing with leading authors from the global scientific community.

Start exploring the collection—download the first chapter of every title for free.

PAPER: Interdisciplinary statistical mechanics

Growth signals determine the topology of evolving networks

Ana Vranić and Marija Mitrović Dankulov*

Institute of Physics Belgrade, University of Belgrade, Pregreva 118, 11080
Belgrade, Serbia

E-mail: anav@ipb.ac.rs and mitrovic@ipb.ac.rs

Received 2 November 2020

Accepted for publication 15 November 2020

Published 22 January 2021



Online at stacks.iop.org/JSTAT/2021/013405
<https://doi.org/10.1088/1742-5468/abd30b>

Abstract. Network science provides an indispensable theoretical framework for studying the structure and function of real complex systems. Different network models are often used for finding the rules that govern their evolution, whereby the correct choice of model details is crucial for obtaining relevant insights. Here, we study how the structure of networks generated with the aging nodes model depends on the properties of the growth signal. We use different fluctuating signals and compare structural dissimilarities of the networks with those obtained with a constant growth signal. We show that networks with power-law degree distributions, which are obtained with time-varying growth signals, are correlated and clustered, while networks obtained with a constant growth signal are not. Indeed, the properties of the growth signal significantly determine the topology of the obtained networks and thus ought to be considered prominently in models of complex systems.

Keywords: random graphs, networks, network dynamics, stochastic processes

 Supplementary material for this article is available [online](#)

J. Stat. Mech. (2021) 013405

Contents

1. Introduction	2
2. Growth signals	3

*Author to whom any correspondence should be addressed.

3. Model of aging nodes with time-varying growth	6
4. Structural differences between networks generated with different growth signals	7
5. Discussion and conclusions	12
Acknowledgments	13
References	14

1. Introduction

Emergent collective behavior is an indispensable property of complex systems [1]. It occurs as a consequence of interactions between a large number of units that compose a complex system, and it cannot be easily predicted from the knowledge about the behavior of these units. The previous research offers definite proof that the interaction network structure is inextricably associated with the dynamics and function of the complex system [2–9]. The structure of complex networks is essential for understanding the evolution and function of various complex systems [10–13].

The structure and dynamics of real complex systems are studied using complex network theory [1, 10, 11]. It was shown that real networks have similar topological properties regardless of their origins [14]. They have broad degree distribution, degree–degree correlations, and power-law scaling of clustering coefficient [11, 14]. Understanding how these properties emerge in complex networks leads to the factors that drive their evolution and shape their structure [2].

The complex network models substantially contribute to our understanding of the connection between the network topology and system dynamics and uncover underlying mechanisms that lead to the emergence of distinctive properties in real complex networks [15–17]. For instance, the famous Barabási–Albert model [15] finds the emergence of broad degree distribution to be a consequence of preferential attachment and network growth. Degree–degree anti-correlations of the internet can be explained, at least to a certain extent, by this constraint [18, 19]. Detailed analysis of the emergence of clustered networks shows that clustering is either the result of finite memory of the nodes [20] or occurs due to triadic closure [21].

Network growth, in combination with linking rules, shapes the network topology [22]. While various rules have been proposed to explain the topology of real networks [10], most models assume a constant rate of network growth, i.e., the addition of a fixed number of nodes at each time step [15, 20, 21]. However, empirical analysis of numerous technological and social systems shows that their growth is time-dependent [23–26]. The time-dependent growth of the number of nodes and links in the networks has been considered as a parameter in uncovering network growth mechanisms [27]. The accelerated growth of nodes in complex networks is the cause of the high heterogeneity in the distribution of web pages among websites [23] and the emergence of highly cited authors in citation networks [26]. The accelerated growth of the number of new links added in each time step changes the shape and scaling exponent of degree distribution

in the Barabási–Albert model [28] and model with preferential attachment with aging nodes [29].

The growth of real systems is not always accelerated. The number of new nodes joining the system varies in time, has trends, and exhibits circadian cycles typical for human behavior [24, 25, 30]. These signals are multifractal and have long-range correlations [31]. Some preliminary evidence shows that the time-varying growth influences the structure and dynamics of the social system and, consequently, the structure of interaction networks in social systems [25, 30, 32–34]. Still, which properties of the real growth signal have the most considerable influence, how different properties influence the topology of the generated networks, and to what extent is an open question.

In this work, we explore the influence of real and computer-generated time-varying growth signals on complex networks' structural properties. We adapt the aging nodes model [35] to enable time-varying growth. We compare the networks' structure using the growing signals from empirical data and randomized signals with ones grown with the constant signal using D -measure [36]. We demonstrate that the growth signal determines the structure of generated networks. The networks grown with time-varying signals have significantly different topology compared to networks generated through constant growth. The most significant difference between topological properties is observed for the values of model parameters for which we obtain networks with broad degree distribution, a common characteristic of real networks [10]. Our results show that real signals, with trends, cycles, and long-range correlations, alter networks' structure more than signals with short-range correlations.

This paper is divided as follows. In section 2, we provide a detailed description of growth signals. In section 3, we briefly describe the original model with aging nodes and structural properties of networks obtained for different values of model parameters [35]. We also describe the changes in the model that we introduce to enable time-varying growth. We describe our results in section 4 and show that the values of D -measure indicate large structural differences between networks grown with fluctuating and ones grown with constant signals. This difference is particularly evident for networks with power-law degree distribution and real growth signals. The networks generated with real signals are correlated and have hierarchical clustering, properties of real networks that do not emerge if we use constant growth. We discuss our results and give a conclusion in section 5.

2. Growth signals

The *growth signal* is the number of new nodes added in each time step. Real complex networks evolve at a different pace, and the dynamics of link creation define the time unit of network evolution. For instance, the co-authorship network grows through establishing a link between two scientists when they publish a paper [37]. In contrast, the links in an online social network are created at a steady pace, often interrupted by sudden bursts [38]. A paper's publication is thus a unit of time for the evolution of co-authorship networks, while the most appropriate time unit for social networks is 1 min or 1 h. While systems may evolve at a different pace, their evolution is often driven by the related mechanisms reflected by the similarity of their structure [10].

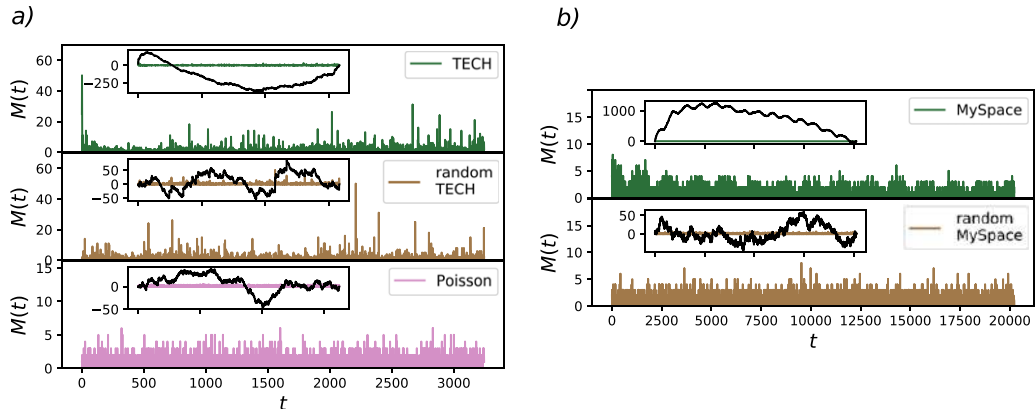


Figure 1. Growth signals for TECH (a) and MySpace (b) social groups, their randomized counterparts, and random signal drawn from Poissonian distribution with mean 1. The cumulative sums of signals' deviations from average mean value are shown in insets.

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 [39] and (b) two months dataset of MySpace social network [40]. TECH is an event-based community where members organize offline events through the Meetup site [39]. 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 [40] with a time resolution of 1 min. The number of newly added nodes for the TECH signal is $N = 3217$, and the length of the signal is $T_s = 3162$ steps. We have shortened the MySpace signal to $T_s = 20\,221$ time steps to obtain the network with $N = 10\,000$ nodes. The signals in the inset of figures 1(a) and (b) show the cumulative sum of deviations of signals from their average mean value, which is 1.017 for TECH and random TECH signal, 0.47 for MySpace and random MySpace, and 1 for Poissonian signal.

Real growth signals have long-range correlations, trends and cycles [25, 30, 40]. We also generate networks using randomized signals and one computer-generated white-noise signal to explore the influence of signals' features on evolving networks' structure. We randomize real signals using a reshuffling procedure. The reshuffling procedure consists of E steps. We randomly select two signal values at two distinct time steps and exchange their position in each step. The number of reshuffling steps is proportional to the length of the signal T_s , and in our case, it equals $100T_s$. Using this procedure, we keep the signal length and mean value, the number of added nodes, and the 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.

We characterize the long-range correlations of the growth signals calculating Hurst exponent [41, 42]. Hurst exponent describes the scaling behavior of time series $M(xt) = x^H M(t)$. It takes values between 0.5 and 1 for long-range correlated signals

and $H = 0.5$ for short-range correlated signals. The most commonly used method for estimating Hurst exponent of real, often non-stationary, temporal signals is detrended fluctuation analysis (DFA) [41]. The DFA removes trends and cycles of real signals and estimates Hurst exponent based on residual fluctuations. The DFA quantifies the scaling behavior of the second-moment fluctuations. However, signals can have deviations in fractal structure with large and small fluctuations that are characterized by different values of Hurst exponents [31].

We use multifractal detrended fluctuation analysis (MFDFA) [31, 43] to estimate multifractal Hurst exponent $H(q)$. For a given time series $\{x_i\}$ with length N , we first define global profile in the form of cumulative sum equation (1), where $\langle x \rangle$ represents an average of the time series:

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

Subtracting the mean of the time series is supposed to eliminate global trends. Insets of figure 1 show global profiles of TECH, MySpace, their randomized signals and Poissonian distribution. The profile of the signal Y is divided into $N_s = \text{int}(N/s)$ 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 $p_{\nu,s}^m$ —polynomial of order m —should be eliminated, and the variance $F^2(\nu, s)$ of detrended signal is calculated as in equation (2):

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^s [Y(j) - p_{\nu,s}^m(j)]^2. \quad (2)$$

Then the q th order fluctuating function is:

$$F_q(s) = \left\{ \frac{1}{2N_s} \sum_{\nu}^{2N_s} [F^2(\nu, s)]^{\frac{q}{2}} \right\}^{\frac{1}{q}}, \quad q \neq 0 \quad (3)$$

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

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.

Figures 1(a) and 2 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. The TECH signal is multifractal, resulting from both broad probability distribution for the values of time series and different long-range correlations of the intervals with small and large fluctuations. Reshuffling of the

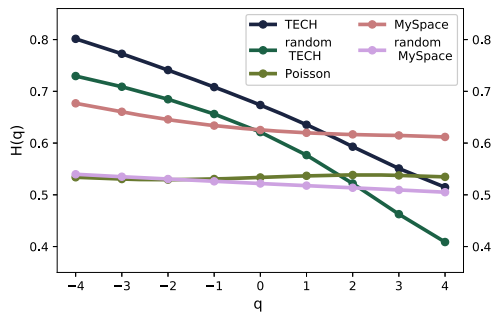


Figure 2. Dependence of Hurst exponent on parameter q for all five signals shown in figure 1 obtained with MF DFA.

time series does not destroy the broad distribution of values, which is the cause for the persistent multifractality of the TECH randomized signal is figure 2.

MySpace signal has a long trend with additional cycles that are a consequence of human circadian rhythm, figure 1(b). Circadian rhythm is an internal process that regulates the sleep-wake cycle and activity, and its period for humans is 24 h [44]. Circadian rhythm leads to periodic changes in online activity during the day and the emergence of a well-defined daily rhythm of activity that we see in figure 1(b). MySpace signal is multifractal for $q < 0$, and has constant value of $H(q)$ for $q > 0$, figure 2. In MF DFA, with negative values of q , we emphasize segments with smaller fluctuations, while for positive q , the emphasis is more on segments with larger fluctuations [43]. Segments with smaller fluctuations have more persistent long-range correlations in both real signals, see figure 2. Randomized MySpace signal and Poissonian signal are monofractal and have short-range with $H = 0.5$ correlations typical for white noise.

Detailed MDFA analysis of real, shuffled, and computer-generated signals are shown in figure S1 and table S1 of the supplementary material (<https://stacks.iop.org/JSTAT/2021/013405/mmedia>). In figure S1 we show in details how the $F_q(s)$ depends on s for different values of parameter q . The curve $F_q(s)$ exhibits different slopes for different values of q for multifractal signals, i.e., TECH, random TECH, and MySpace. $F_q(s)$ curves for monofractal signals are parallel. We provide the estimated values of $H(q)$ with estimated errors for q in a range from -4 to 4 for all five signals in table S1 of the supplementary material.

3. Model of aging nodes with time-varying growth

To study the influence of temporal fluctuations of growth signal on network topology, we need a model with linking rules where linking probability between network nodes depends on time. We use a network model with aging nodes [35]. In this model, the probability of linking the newly added node and the old one is proportional to their age difference and an old node's degree. In the original version of the model, one node is added to the network and linked to one old node in each time step. The old node is

chosen according to probability

$$\Pi_i(t) \sim k_i(t)^\beta \tau_i^\alpha \quad (4)$$

where $k_i(t)$ is a degree of a node i at time t , and τ_i is age difference between node i and newly added node. As was shown in [35], the values of model parameters β and α determine the topological properties of the resulting networks grown with the constant signal. According to this work, the networks generated using constant growth signals are uncorrelated trees for all values of model parameters. The phase diagram in α - β plain, obtained for $\beta > 0$ and $\alpha < 0$, shows that the degree distribution $P(k) \sim k^{-\gamma}$ with $\gamma = 3$ is obtained only along the line $\beta(\alpha^*)$, see [35] and figure S2 in the supplementary material. For $\alpha > \alpha^*$ networks have gel-like small world behavior, while for $\alpha < \alpha^*$ but close to line $\beta(\alpha^*)$ networks have stretched exponential shape of degree distribution [35].

Here we slightly change the original aging model [35] to enable the addition of more than one node and more than one link per newly added node in each time step. In each time step, we add $M \geq 1$ new nodes to the network and link them to $L \geq 1$ old nodes according to probability Π_i given in equation (4). Again, the networks with broad degree distribution are only generated for the combination of the model parameters along the critical line $\beta(\alpha^*)$. This line's position in the α - β plane changes with link density, while the addition of more than one node in each time step does not influence its position. Our analysis shows that the critical line's position is independent of the growth signal's properties, see figure S2 in the supplementary material showing phase diagram. For instance, for $L = 1$ networks and $\alpha = -1.25$ and $\beta = 1.5$ we obtain networks with power-law degree, while for $L = 2$ and $\beta = 1.5$ we need to increase the value of parameter α to -1.0 in order to obtain networks with broad degree distribution. Networks obtained for the values of model parameters $\beta(\alpha^*)$, $L \geq 2$, and constant growth have power-law degree distribution, are uncorrelated and have a finite non-zero value of clustering coefficient which does not depend on node degree, figure 4(b). If we fix the value of parameter β and lower down the value of parameter α to -1.5 , the resulting networks are uncorrelated with a small value of clustering coefficient, see figure 4(a). For $\alpha < \alpha^*$ we obtain networks with stretched exponential degree distribution, without degree-degree correlations and small value of clustering exponent that does not depend on node degree (see figure S2 in the supplementary material). For $\alpha \ll \alpha^*$ the resulting networks are regular graphs. If we keep the value of α to 1.0 but increase the value β to 2.0 we enter the region of small world gels, see figure 4(c). The networks created for the values of $\alpha > \alpha^*$ are correlated networks with power-law dependence of the clustering coefficient on the degree (see figure S2 in the supplementary material). However, these networks do not have a power-law degree distribution.

The master equation approach is useful for studying the model with aging nodes when $M(t) = 1$ [45]. However, this approach is not sufficient for time-varying growth signals. 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 $-\infty < \alpha \leq 0$ and $\beta \geq 1$ and constant L .

4. Structural differences between networks generated with different growth signals

We generate networks for different values of L , and different growth signal profiles $M(t)$. To examine how these properties influence the network structure, we compare the network structure obtained with different growth signals with networks of the same size grown with constant signal $M = 1$. The $M = 1$ is the closest constant value to average values of the signals, which are 1.017 for TECH, 0.47 for MySpace, and 1 for Poissonian signals. We explore the parameter space of the model by generating networks for pairs of values (α, β) in the range $-3 \leq \alpha \leq -0.5$ and $1 \leq \beta \leq 3$ with steps 0.5. For each pair of (α, β) we generated networks of different link density by varying parameter $L \in 1, 2, 3$, and for each combination of (α, β, L) , we generate a sample of 100 networks and compare the structure of the networks grown with $M = 1$ with the ones grown with $M(t)$ shown in figure 1.

We quantify topological differences between two networks using D -measure defined in [36]

$$D(G, G') = \omega \left| \sqrt{\frac{J(P_1, \dots, P_N)}{\log(d+1)}} - \sqrt{\frac{J(P'_1, \dots, P'_N)}{\log(d'+1)}} \right| + (1 - \omega) \sqrt{\frac{J(\mu_G, \mu_{G'})}{\log 2}}. \quad (5)$$

D -measure captures the topological differences between two networks, G and G' , on a local and global level. The first term in equation (5) evaluates dissimilarity between two networks on a local level. For each node in the network G one can define the distance distribution $P_i = \{p_i(j)\}$, where $p_i(j)$ is a fraction of nodes in network G that are connected to node i at distance j . The set of N node-distance distributions $\{P_1, \dots, P_N\}$ contains a detailed information about network's topology. The heterogeneity of a graph G in terms of connectivity distances is measured through node network dispersion (NND). In [36] authors estimate NND as Jensen–Shannon divergence between N distance distributions $J(P_1, \dots, P_N)$ normalized by $\log(d+1)$, where d is diameter of network G , and show that NND captures relevant features of heterogeneous networks. The difference between NNDs for graph G and G' captures the dissimilarity between the graph's connectivity distance profile.

However, certain graphs, such as k -regular graphs, have $\text{NND} = 0$ and can not be compared using NND. For these reasons, authors also introduce average node distance distribution of a graph $\mu(G) = \{\mu(1), \dots, \mu(d)\}$, where $\mu(k)$ is the fraction of all pair of nodes in the network G that are at a distance k . The Jensen–Shannon divergence between $\mu(G)$ and $\mu(G')$ measures the difference between nodes' average connectivity in a graph G and G' . This term captures the differences between nodes on a global scale.

The original definition of D -measure also includes the third term, which quantifies dissimilarity in node α -centrality. The term can be omitted without precision loss [36]. The parameter ω in equation (5) determines the weight of each term. The extensive analysis shows that the choice $\omega = 0.5$ is the most appropriate for quantifying structural differences between two networks [36].

The D -measure takes the value between 0 and 1. The lower the value of D -measure is the more similar two networks are, with $D = 0$ for isomorphic graphs. The D -measure

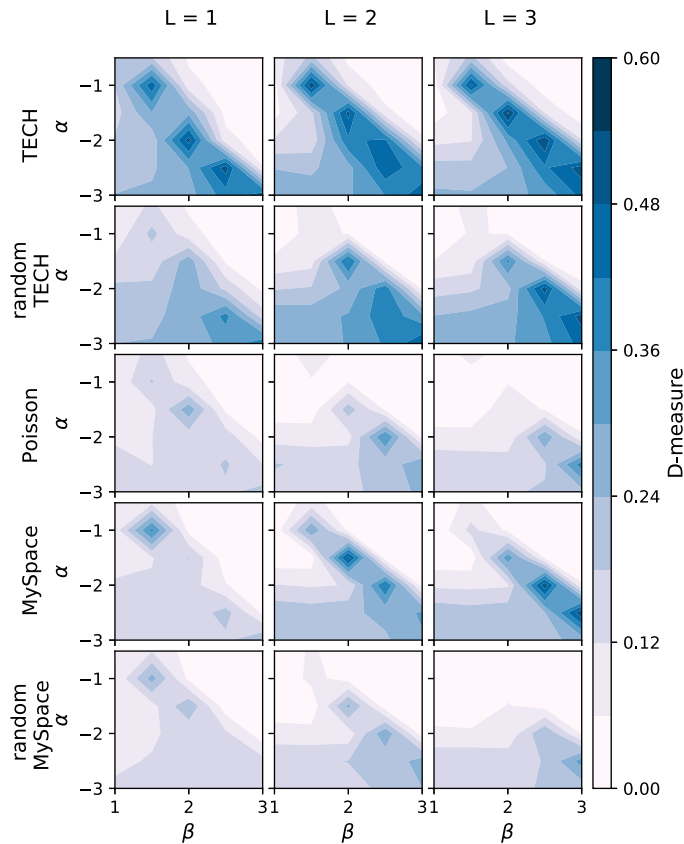


Figure 3. The comparison of networks grown with growth signals shown in figure 1 versus ones grown with constant signal $M = 1$, for value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in each time step.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

outperforms previously used network dissimilarity measures such as Hamming distance and graph editing distance and clearly distinguishes between networks generated with the same model but with different values of model parameters [36].

For each pair of networks, one grown with constant and one with the fluctuating signal, we calculate the D -measure. The structural difference between networks grown with constant and fluctuating growth signal for fixed L and values of parameters α and β is obtained by averaging the D -measure calculated between all possible pairs of networks, see figure 3. We observe the non-zero value of D -measure for all time-varying signals. The D -measure has the largest value in the region around the line $\beta(\alpha^*)$. The values of D -measure in this region are similar to ones observed when comparing Erdős–Rényi graphs grown with linking probability below and above critical value [36]. For values $\beta < \beta(\alpha^*)$, the structural differences between networks grown with constant signal and $M(t)$ still exist, but they become smaller as we are moving away from the critical line. Networks obtained with constant signal and fluctuating signals have statistically similar structural properties in the region of small-world network gels, i.e., $\alpha > \alpha^*$.

We focus on the region around the critical line and observe the significant structural discrepancies between networks created for constant versus time-dependent growth signals for all signals regardless of their features. However, the value of D -measure depends on the signal's properties, figure 3. Networks grown with multifractal signals, TECH, random TECH, and MySpace signals, are the most different from those created by a constant signal. The D -measure has the maximum value for the original TECH signal, with $D_{\max} = 0.552$, the signal with the most pronounced multifractal properties among all signals shown in figure 2. Networks generated with randomized MySpace signal and Poisson signal are the least, but still notably dissimilar from those created with $M = 1$.

Randomized MySpace signal and Poissonian signal are monofractal signals with Hurst exponent $H = 0.5$. To investigate the influence of monofractal correlated signals on the network structure, we generate six signals with a different value of $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, see figure S3 in the supplementary material. We use each of these signals to generate networks following the same procedure as for signals shown in figure 1. The results shown in figure S4 of the supplementary material confirm that short-range correlated signals create networks with different structures from ones grown with the constant signal. The increase of the Hurst exponent leads to increases in the D -measure. However, D -measure's maximal value is smaller than one observed for multifractal signals shown in figure 3.

The value of D -measure rises with a decline of α^* . This observation can be explained by examining linking rules and how model parameters determine linking dynamics between nodes. The ability of a node to acquire a link declines with its age and grows with its degree. A node's potential to become a hub, node with a degree significantly larger than average network degree, depends on the number of nodes added to the network in the T time steps after its birth. The length of the interval T decreases with parameter α . For constant signal, the number of nodes added during this time interval is constant and equal to MT . For fluctuating growth signals, the number of added nodes during the time T varies with time. In signals that have a broad distribution of fluctuations, like TECH signals, the peaks of the number of newly added nodes lead to the emergence of one or several hubs and super hubs. The emergence of super hubs, nodes connected to more than 30% of the nodes in the network, significantly alters the network's topology. For instance, super hubs' existence lowers the value of average path length and network diameter [10]. The emergence of hubs occurs for values of parameter α relative close to -1.0 for signals with long-range correlations. As we decrease the parameter α , the fluctuations present in the time-varying signals become more important, and we observe the emergence of hubs even for the white-noise signals. The trends present in real growth signals further promote the emergence of hubs. The impact of fluctuations and their temporal features on the structure of complex networks increases with link density.

The large number of structural properties observed in real networks are often consequences of particular degree distributions, degree correlations, and clustering coefficient [47]. Figure 4 shows the degree distribution $P(k)$, dependence of average neighboring degree on node degree $\langle k \rangle_{nn}(k)$, and dependence of clustering coefficient on node degree $c(k)$ for networks with average number of links per node $L = 2$. The significant structural differences between networks grown with real time-varying and constant signals

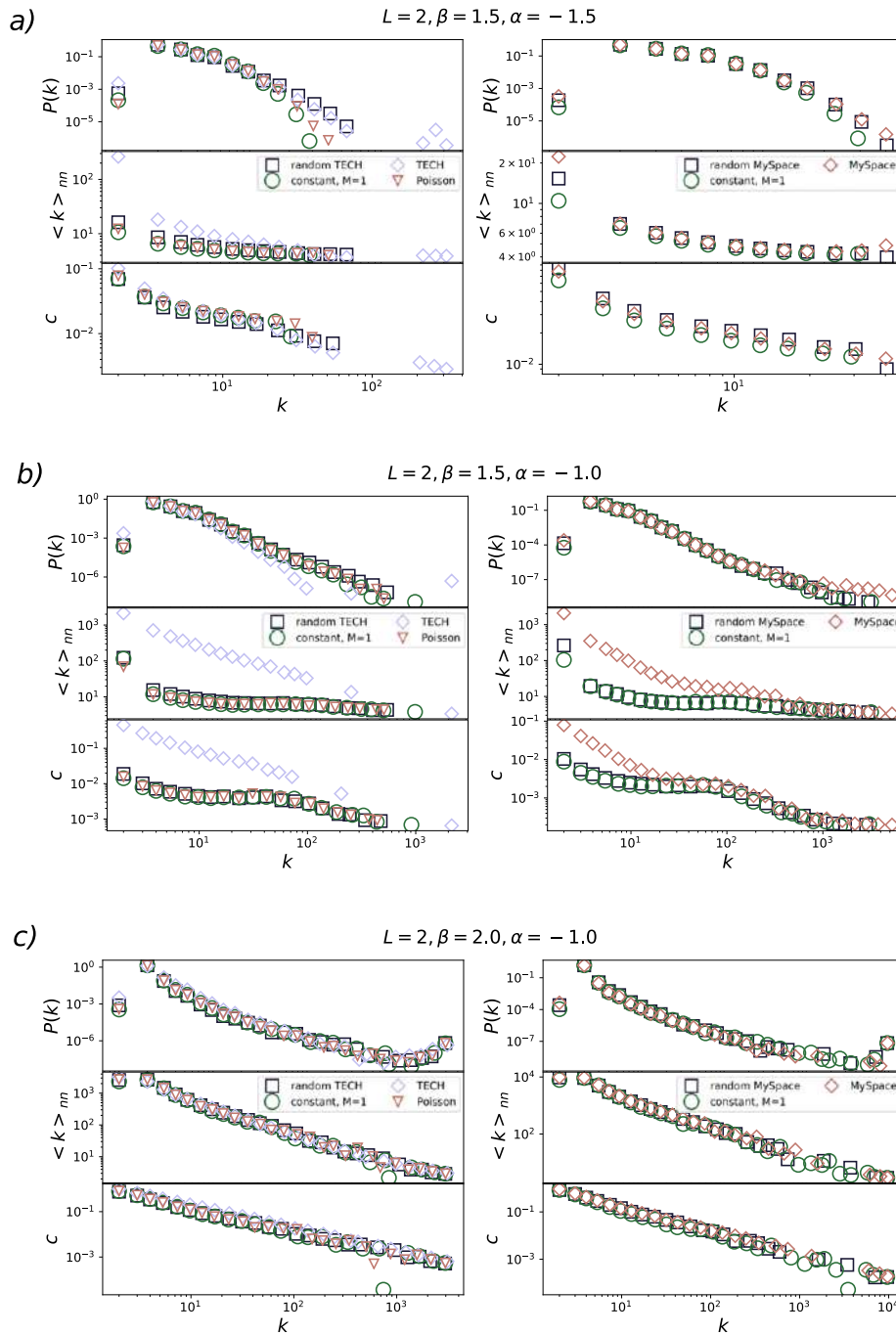


Figure 4. Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have the values $\alpha = -1.5$, $\beta = 1.5$ (a), $\alpha = -1.0$, $\beta = 1.5$ (b), $\alpha = -1.0$, $\beta = 2.0$ (c), and $L = 2$ for all networks.

are observed for the values of model parameters $\alpha = -1.0$ and $\beta = 1.5$, figures 3 and 4(b). The degree distribution of networks generated for real signals shows the occurrence of super hubs in these networks. In contrast, degree distributions of networks generated with white-noise like signals do not differ from one created with constant signal, figure 4(b). Networks obtained for the real signals are disassortative and have a hierarchical structure, i.e., their clustering coefficient decreases with the degree. On the other hand, networks generated with constant and randomized signals are uncorrelated, and their clustering weakly depends on the degree.

We observe a much smaller, but still noticeable, difference between the topological properties of networks evolved with constant and time-varying signal for $\alpha < \alpha^*$, figure 4(a). The difference is particularly observable for degree distribution and dependence of average neighboring degree on node degree of networks grown with real TECH signal. The fluctuations of time-varying growth signals do not influence the topological properties of small-world gel networks, figure 4(c). For $\alpha > \alpha^*$, the super hubs emerge even with the constant growth. Since this is the mechanism through which the fluctuations alter the structure of evolving networks for $\alpha \leq \alpha^*$, the features of the growth signals cease to be relevant.

5. Discussion and conclusions

We demonstrate that the resulting networks' structure depends on the time-varying signal features that drive their growth. The previous research [25, 30] indicated the possible influence of temporal fluctuations on network properties. Our results show that growth signals' temporal properties 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 [35].

We observe the most substantial dissimilarity in network structure along the critical line, the values of model parameters for which we generate broad degree distribution networks. Figure 3 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 the 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 Erdős–Rényi graphs [36] indicating the considerable structural difference between these networks. Our findings are further confirmed in figure 4(b). The networks generated with signals with 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^* .

Randomized and computer-generated signals do not have trends or cycles. Nevertheless, networks grown with these signals have a significantly different structure from ones grown with constant M . Our results demonstrate that growth signals' temporal

fluctuations are the leading cause for the structural differences between networks evolved with the constant and time-varying signal. We observe the smallest, but significant, difference between networks generated with constant M and monofractal signal with short-range correlations. As we increase the Hurst exponent, the value of the D -measure increases. The most considerable differences are observed for multifractal signals TECH, random TECH, and MySpace.

The value of D -measure declines as we move away from the critical line, figure 3. The primary mechanism through which the fluctuations influence the structure of evolved networks is the emergence of hubs and super hubs. For values of $\alpha \ll \alpha^*$, the nodes attach to their immediate predecessors creating regular networks without hubs. For $\alpha \lesssim \alpha^*$ 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 4(a). By definition, small-world networks generated for $\alpha > \alpha^*$ have super-hubs [35] regardless of the growth signal. 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 α - β plane.

In this work, we focus on the role of the node growth signal in evolving networks' structure. However, real networks do not evolve only due to the addition of new nodes, but also through addition of new links [27–29, 38]. Furthermore, the deactivation of nodes [48] and the links [48] influence the evolving networks' structure. Each of these processes alone can result in a different network despite having the same linking rules. The next step would be to examine how different combinations of these processes influence the evolving networks' structure. For instance, in [28], authors have examined the influence of the time-dependent number of added links $L(t)$ on the Barabási–Albert networks' structure. They show that as long as the average value of time-dependent signal $\langle L(t) \rangle$ is independent of time, the generated networks have a similar structure as Barabási–Albert networks, and that the degree distribution depends strongly on the behavior of $\langle L(t) \rangle$. It would be interesting to examine how correlated $L(t)$ signals influence networks' structure with aging nodes, where the age of a node plays a vital role in linking between new and old nodes. Moreover, we expect that the combination of time-varying growth of the number of nodes and the number of links will significantly influence these networks' structure.

Evolving network models are an essential tool for understanding the evolution of social, biological, and technological networks and mechanisms that drive it [10]. The most common assumption is that these networks evolve by adding a fixed number of nodes in each time step [10]. So far, the focus on developing growing network models was on linking rules and how different rules lead to networks of various structural properties [10]. Growth signals of real systems are not constant [25, 30]. They are multifractal, characterised with long-range correlations [25], trends and cycles [40]. 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 [12]. Our results imply that modeling of social and technological networks should also include non-constant growth. Its combination with local linking rules can significantly alter the structure of generated networks.

Acknowledgments

We acknowledge funding provided by the Institute of Physics Belgrade, through the grant by the Ministry of Education, Science, and Technological Development of the Republic of Serbia. This research was supported by the Science Fund of the Republic of Serbia, 65241005, AI-ATLAS. Numerical simulations were run on the PARADOX-IV supercomputing facility at the Scientific Computing Laboratory, National Center of Excellence for the Study of Complex Systems, Institute of Physics Belgrade. The work of MMD was, in part, supported by the Ito Foundation fellowship.

References

- [1] Ladyman J, Lambert J and Wiesner K 2013 What is a complex system? *Euro J. Phil. Sci.* **3** 33
- [2] Barrat A, Barthélemy M and Vespignani A 2008 *Dynamical Processes on Complex Networks* (Cambridge: Cambridge University Press)
- [3] Pascual M *et al* 2006 *Ecological Networks: Linking Structure to Dynamics in Food Webs* (Oxford: Oxford University Press)
- [4] Castellano C, Fortunato S and Loreto V 2009 Statistical physics of social dynamics *Rev. Mod. Phys.* **81** 591
- [5] Gosak G, Markovič R, Dolensek J, Rupnik M S, Marhl M, Stožer A and Perc M 2018 Network science of biological systems at different scales: a review *Phys. Life Rev.* **24** 118–35
- [6] Arenas A, Díaz-Guilera A, Kurths J, Moreno Y and Zhou C 2008 Synchronization in complex networks *Phys. Rep.* **469** 93
- [7] Boccaletti S, Almendral J A, Guan S, Leyva I, Liu Z, Sendiña-Nadal I, Wang Z and Zou Y 2016 Explosive transitions in complex networks' structure and dynamics: percolation and synchronization *Phys. Rep.* **660** 1
- [8] Chen H, Zhang H and Shen C 2018 Double phase transition of the Ising model in core-periphery networks *J. Stat. Mech.* **063402**
- [9] Kuga K and Tanimoto J 2018 Impact of imperfect vaccination and defense against contagion on vaccination behavior in complex networks *J. Stat. Mech.* **113402**
- [10] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175
- [11] Newman M E J 2010 *Networks: An Introduction* (Oxford: Oxford University Press)
- [12] Holme P and Saramäki J 2012 Temporal networks *Phys. Rep.* **519** 97
- [13] Boccaletti S, Bianconi G, Criado R, Del Genio C I, Gómez-Gardeñes J, Romance M, Sendiña-Nadal I, Wang Z and Zanin M 2014 The structure and dynamics of multilayer networks *Phys. Rep.* **544** 1
- [14] Barabási A-L 2009 Scale-free networks: a decade and beyond *Science* **325** 412
- [15] Barabási A-L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509
- [16] Tadić B 2001 Dynamics of directed graphs: the world-wide web *Physica A* **293** 273
- [17] Mitrović M and Tadić B 2009 Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities *Phys. Rev. E* **80** 026123
- [18] Maslov S, Sneppen K and Zaliznyak A 2004 Detection of topological patterns in complex networks: correlation profile of the internet *Physica A* **333** 529
- [19] Park J and Newman M E J 2003 Origin of degree correlations in the internet and other networks *Phys. Rev. E* **68** 026112
- [20] Klemm K and Eguiluz V M 2002 Highly clustered scale-free networks *Phys. Rev. E* **65** 036123
- [21] Serrano M A and Boguná M 2005 Tuning clustering in random networks with arbitrary degree distributions *Phys. Rev. E* **72** 036133
- [22] Vázquez A 2003 Growing network with local rules: preferential attachment, clustering hierarchy, and degree correlations *Phys. Rev. E* **67** 056104
- [23] Huberman B A and Adamic L A 1999 Growth dynamics of the world-wide web *Nature* **401** 131
- [24] Mitrović M and Tadić B 2010 Bloggers behavior and emergent communities in blog space *Eur. Phys. J. B* **73** 293
- [25] Dankulov M M, Melnik R and Tadić B 2015 The dynamics of meaningful social interactions and the emergence of collective knowledge *Sci. Rep.* **5** 1

- [26] Liu J, Li J, Chen Y, Chen X, Zhou Z, Yang Z and Zhang C-J 2019 Modeling complex networks with accelerating growth and aging effect *Phys. Lett. A* **383** 1396
- [27] Pham T, Sheridan P and Shimodaira H 2016 Joint estimation of preferential attachment and node fitness in growing complex networks *Sci. Rep.* **6** 32558
- [28] Sen P 2004 Accelerated growth in outgoing links in evolving networks: deterministic versus stochastic picture *Phys. Rev. E* **69** 046107
- [29] Dorogovtsev S N and Mendes J F F 2001 Effect of the accelerating growth of communications networks on their structure *Phys. Rev. E* **63** 025101
- [30] Mitrović M and Tadić B 2012 *Emergence and Structure of Cybercommunities (Springer Optimization and Its Applications)* vol 57 (Berlin: Springer) p 209
- [31] Kantelhardt J W, Zschiegner S A, Koscielny-Bunde E, Havlin S, Bunde A and Stanley H E 2002 Multifractal detrended fluctuation analysis of nonstationary time series *Physica A* **316** 87
- [32] Mitrović M, Paltoglou G and Tadić B 2011 Quantitative analysis of bloggers' collective behavior powered by emotions *J. Stat. Mech.* **P02005**
- [33] Tadić B, Dankulov M M and Melnik R 2017 Mechanisms of self-organized criticality in social processes of knowledge creation *Phys. Rev. E* **96** 032307
- [34] Tadić B and Šuvakov M 2013 Can human-like bots control collective mood: agent-based simulations of online chats *J. Stat. Mech.* **P10014**
- [35] Hajra K B and Sen P 2004 Phase transitions in an aging network *Phys. Rev. E* **70** 056103
- [36] Schieber T A, Carpi L, Díaz-Guilera A, Pardalos P M, Masoller C and Ravetti M 2017 Quantification of network structural dissimilarities *Nat. Commun.* **8** 1
- [37] Sarigöl E, Pfitzner R, Scholtes I, Garas A and Schweitzer F 2014 Predicting scientific success based on coauthorship networks *EPJ Data Sci.* **3** 9
- [38] Myers S A and Leskovec J 2014 The bursty dynamics of the twitter information network *Proc. 23rd Int. Conf. on World Wide Web* 913
- [39] Smiljanić J and Dankulov M M 2017 Associative nature of event participation dynamics: a network theory approach *PloS One* **12** e0171565
- [40] Šuvakov M, Mitrović M, Gligorić V and Tadić B 2013 How the online social networks are used: dialogues-based structure of MySpace *J. R. Soc. Interface* **10** 20120819
- [41] Peng C-K, Buldyrev S V, Havlin S, Simons M, Stanley H E and Goldberger A L 1994 Mosaic organization of DNA nucleotides *Phys. Rev. E* **49** 1685
- [42] Kantelhardt J W, Koscielny-Bunde E, Rego H H A, Havlin S and Bunde A 2001 Detecting long-range correlations with detrended fluctuation analysis *Physica A* **295** 441
- [43] Fürst EAFI Ihlen E A 2012 Introduction to multifractal detrended fluctuation analysis in Matlab *Front. Physiol.* **3** 141
- [44] Wever R A 2013 *The Circadian System of Man: Results of Experiments under Temporal Isolation* (Berlin: Springer)
- [45] Dorogovtsev S N and Mendes J F F 2001 Scaling properties of scale-free evolving networks: continuous approach *Phys. Rev. E* **63** 056125
- [46] Orsini C *et al* 2015 Quantifying randomness in real networks *Nat. Commun.* **6** 8627
- [47] Tian L, Zhu C-P, Shi D-N, Gu Z-M and Zhou T 2006 Universal scaling behavior of clustering coefficient induced by deactivation mechanism *Phys. Rev. E* **74** 046103
- [48] Gagen M J and Mattick J S 2005 Accelerating, hyperaccelerating, and decelerating networks *Phys. Rev. E* **72** 016123

Charge transport in the Hubbard model at high temperatures: Triangular versus square lattice

A. Vranić¹, J. Vučičević¹, J. Kokalj^{2,3}, J. Skolimowski^{3,4}, R. Žitko^{3,5}, J. Mravlje³, and D. Tanasković¹


¹*Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia*

²*University of Ljubljana, Faculty of Civil and Geodetic Engineering, Jamova 2, Ljubljana, Slovenia*

³*Jožef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia*

⁴*International School for Advanced Studies (SISSA), Via Bonomea 265, I-34136 Trieste, Italy*

⁵*University of Ljubljana, Faculty of Mathematics and Physics, Jadranska 19, Ljubljana, Slovenia*

 (Received 3 June 2020; revised 7 August 2020; accepted 2 September 2020; published 21 September 2020)

High-temperature bad-metal transport has been recently studied both theoretically and in experiments as one of the key signatures of strong electronic correlations. Here we use the dynamical mean field theory and its cluster extensions, as well as the finite-temperature Lanczos method to explore the influence of lattice frustration on the thermodynamic and transport properties of the Hubbard model at high temperatures. We consider the triangular and the square lattices at half-filling and at 15% hole doping. We find that for $T \gtrsim 1.5t$ the self-energy becomes practically local, while the finite-size effects become small at lattice size 4×4 for both lattice types and doping levels. The vertex corrections to optical conductivity, which are significant on the square lattice even at high temperatures, contribute less on the triangular lattice. We find approximately linear temperature dependence of dc resistivity in doped Mott insulator for both types of lattices.

DOI: [10.1103/PhysRevB.102.115142](https://doi.org/10.1103/PhysRevB.102.115142)

I. INTRODUCTION

Strong correlation effects in the proximity of the Mott metal-insulator transition are among the most studied problems in modern condensed matter physics. At low temperatures, material-specific details play a role, and competing mechanisms can lead to various types of magnetic and charge density wave order, or superconductivity [1–5]. At higher temperatures, physical properties become more universal, often featuring peculiarly high and linear-in-temperature resistivity (the bad-metal regime) [6–12] and gradual metal-insulator crossover obeying typical quantum critical scaling laws [13–17].

There are a number of theoretical studies of transport in the high- T regime based on numerical solutions of the Hubbard model [10,12,13,18,19], high- T expansion [20], and field theory [21–23]. Finding numerically precise results is particularly timely having in mind a very recent laboratory realization of the Hubbard model using ultracold atoms on the optical lattice [24]. This system enables fine tuning of physical parameters in a system without disorder and other complications of bulk crystals, which enables a direct comparison between theory and experiment. In our previous work (Ref. [25]) we have performed a detailed analysis of single- and two-particle correlation functions and finite-size effects on the square lattice using several complementary state-of-the-art numerical methods, and established that a finite-temperature Lanczos method (FTLM) solution on the 4×4 lattice is nearly exact at high temperatures. The FTLM, which calculates the correlation functions directly on the real-frequency axis, is recognized [25] as the most reliable method for calculating the transport properties of the Hubbard model at high temperatures. The dependence of charge transport and

thermodynamics on the lattice geometry has not been examined in Ref. [25] and it is the subject of this work.

Numerical methods that we use are (cluster) dynamical mean field theory (DMFT) and FTLM. The DMFT treats an embedded cluster in a self-consistently determined environment [26]. Such a method captures long-distance quantum fluctuations, but only local (in single-site DMFT), or short-range correlations (in cluster DMFT) [27]. The results are expected to converge faster with the size of the cluster than in the FTLM, which treats a finite cluster with periodic boundary conditions [28]. FTLM suffers from the finite-size effects in propagators as well as in correlations. The conductivity calculation in DMFT is, however, restricted just to the bubble diagram, while neglecting the vertex corrections. Approximate calculation of vertex corrections is presented in few recent works [29–34]. This shortcoming of DMFT is overcome in FTLM where one calculates directly the current-current correlation function which includes all contributions to the conductivity. Also, the FTLM calculates conductivity directly on the real-frequency axis, thus eliminating the need for analytical continuation from the Matsubara axis which can, otherwise, lead to unreliable results (see Supplemental Material of Ref. [25]). Both DMFT and FTLM methods are expected to work better at high temperatures [35] when single- and two-particle correlations become more local, and finite-size effects less pronounced. Earlier work has shown that the single-particle nonlocal correlations become small for $T \gtrsim t$ for both the triangular and the square lattices [25,36,37].

In this paper we calculate the kinetic and potential energy, specific heat, charge susceptibility, optical and dc conductivity in the Hubbard model on a triangular lattice and make a comparison with the square-lattice results. We consider strongly correlated regime at half-filling and at 15% hole doping. In

agreement with the expectations, we find that at high temperatures, $T \gtrsim 1.5t$, the nonlocal correlations become negligible and the results for thermodynamic quantities obtained with different methods coincide, regardless of the lattice type and doping. At intermediate temperatures, $0.5t \lesssim T \lesssim 1.5t$, the difference between DMFT and FTLM remains rather small. Interestingly, we do not find that the thermodynamic quantities are more affected by nonlocal correlations on the square lattice in this temperature range, although the self-energy becomes more local on the triangular lattice due to the magnetic frustration. On the other hand, the vertex corrections to optical conductivity remain important even at high temperatures for both lattice types, but we find that they are substantially smaller in the case of a triangular lattice. For the doped triangular and square lattice the temperature dependence of resistivity is approximately linear for temperatures where the finite-size effects become negligible and where the FTLM solution is close to exact.

The paper is organized as follows. In Sec. II we briefly describe different methods for solving the Hubbard model. Thermodynamic and charge transport results are shown in Sec. III, and conclusions in Sec. IV. The Appendix contains a detailed comparison of the DMFT optical conductivity obtained with different impurity solvers, a brief discussion of the finite-size effects at low temperatures, and an illustration of the density of states in different transport regimes.

II. MODEL AND METHODS

We consider the Hubbard model given by the Hamiltonian

$$H = -t \sum_{\langle i,j \rangle, \sigma} c_{i\sigma}^\dagger c_{j\sigma} + U \sum_i n_{i\uparrow} n_{i\downarrow} - \mu \sum_{i\sigma} n_{i\sigma}, \quad (1)$$

where t is the hopping between the nearest neighbors on either triangular or square lattice. $c_{i\sigma}^\dagger$ and $c_{i\sigma}$ are the creation and annihilation operators, U is the onsite repulsion, $n_{i\sigma}$ is the occupation number operator, and μ is the chemical potential. We set $U = 10t$, $t = 1$, lattice constant $a = 1$, $e = \hbar = k_B = 1$ and consider the paramagnetic solution for $p = 1 - n = 1 - \sum_{\sigma} n_{\sigma} = 0.15$ hole doping and at half-filling.

We use the FTLM and DMFT with its cluster extensions to solve the Hamiltonian. FTLM is a method based on the exact diagonalization of small clusters (4×4 in this work). It employs the Lanczos procedure to obtain approximate eigenstates and uses sampling over random starting vectors to calculate the finite-temperature properties from the standard expectation values [28]. To reduce the finite-size effects, we further employ averaging over twisted boundary conditions.

The (cluster) DMFT equations reduce to solving a (cluster) impurity problem in a self-consistently determined effective medium. We consider the single-site DMFT, as well as two implementations of cluster DMFT: cellular DMFT (CDMFT) [38,39] and dynamical cluster approximation (DCA) [27]. In DMFT the density of states is the only lattice-specific quantity that enters into the equations. In CDMFT we construct the supercells in the real space and the self-energy obtains short-ranged nonlocal components within the supercell. In DCA we divide the Brillouin zone into several patches and the number of independent components of the self-energy equals the number of inequivalent patches. The DCA results on 4×4 and

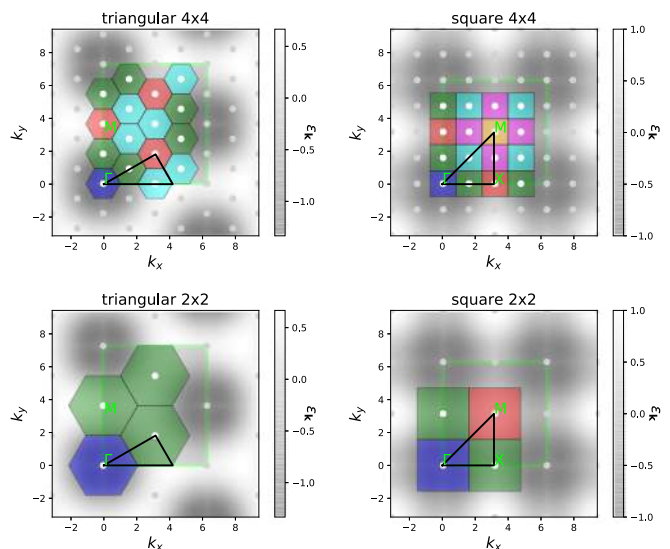


FIG. 1. DCA patches in the Brillouin zone. The irreducible Brillouin zone is marked by the black triangle. The dispersion relation is shown in gray shading. Note the position of the Γ point in the center of the first Brillouin zone which is not marked in this figure.

2×2 clusters are obtained by patching the Brillouin zone in a way that obeys the symmetry of the lattice, as shown in Fig. 1. As the impurity solver we use the continuous-time interaction expansion (CTINT) quantum Monte Carlo (QMC) algorithm [40,41]. In the single-site DMFT we also use the numerical renormalization group (NRG) impurity solver [42–45].

The (cluster) DMFT with QMC impurity solver (DMFT-QMC) gives the correlation functions on the imaginary (Matsubara) frequency axis, from which static quantities can be easily evaluated. The kinetic energy per lattice site is equal to

$$E_{\text{kin}} = \frac{1}{N} \sum_{\mathbf{k}} \varepsilon_{\mathbf{k}} n_{\mathbf{k}\sigma} = \frac{2}{N} \sum_{\mathbf{k}} \varepsilon_{\mathbf{k}} G_{\mathbf{k}}(\tau = 0^-), \quad (2)$$

where for the triangular lattice $\varepsilon_{\mathbf{k}} = -2t[\cos k_x + 2 \cos(\frac{1}{2}k_x) \cos(\frac{\sqrt{3}}{2}k_y)]$ and for the square lattice $\varepsilon_{\mathbf{k}} = -2t(\cos k_x + \cos k_y)$ (gray shading in Fig. 1). The noninteracting band for the triangular lattice goes from $-6t$ to $3t$ with the van Hove singularity at $\varepsilon = t$. The potential energy is equal to

$$E_{\text{pot}} = Ud = \frac{1}{N} T \sum_{\mathbf{k}, i\omega_n} e^{i\omega_n 0^+} G_{\mathbf{k}}(i\omega_n) \Sigma_{\mathbf{k}}(i\omega_n), \quad (3)$$

where $d = \langle n_{i\uparrow} n_{i\downarrow} \rangle$ is the average double occupation. In DCA the cluster double occupation is the same as on the lattice, and we used the direct calculation of d in the cluster solver to cross check the consistency and precision of the numerical data. In CDMFT we calculated E_{pot} from periodized quantities G and Σ , where the periodization is performed on the self-energy and then the lattice Green's function is calculated from it. The total energy is $E_{\text{tot}} = E_{\text{kin}} + E_{\text{pot}}$. The specific heat $C = dE_{\text{tot}}/dT|_n$ is obtained by interpolating $E_{\text{tot}}(T)$ and then taking a derivative with respect to temperature. C is shown only in the DMFT solution where we had enough points

at low temperatures. The charge susceptibility $\chi_c = \partial n / \partial \mu$ is obtained from a finite difference using two independent calculations with μ that differs by a small shift $\delta\mu = 0.1t$. In the FTLM, C and χ_c are calculated without taking the explicit numerical derivative since the derivation can be done analytically from a definition of the expectation values,

$$\begin{aligned} C &= C_\mu - \frac{T\zeta^2}{\chi_c} \\ &= \frac{1}{N} \frac{1}{T^2} \left[\langle H^2 \rangle - \langle H \rangle^2 - \frac{(\langle HN_e \rangle - \langle H \rangle \langle N_e \rangle)^2}{\langle N_e^2 \rangle - \langle N_e \rangle^2} \right], \end{aligned} \quad (4)$$

which is directly calculated in FTLM. Here, $C_\mu = \frac{1}{N} \frac{1}{T^2} [\langle (H - \mu N_e)^2 \rangle - \langle H - \mu N_e \rangle^2]$, $\zeta = \frac{1}{N^2} \frac{1}{T^2} [\langle (H - \mu N_e) N_e \rangle - \langle H - \mu N_e \rangle \langle N_e \rangle]$, $\chi_c = \frac{1}{N} \frac{1}{T} (\langle N_e^2 \rangle - \langle N_e \rangle^2)$, and $N_e = \sum_{i\sigma} n_{i\sigma}$ is the operator for the total number of electrons on the lattice.

We calculate the conductivity using DMFT and FTLM. Within the DMFT the optical conductivity is calculated from the bubble diagram as

$$\begin{aligned} \sigma(\omega) &= \sigma_0 \iint d\varepsilon d\nu X(\varepsilon) A(\varepsilon, \nu) A(\varepsilon, \nu + \omega) \\ &\quad \times \frac{f(\nu) - f(\nu + \omega)}{\omega}, \end{aligned} \quad (5)$$

where $X(\varepsilon) = \frac{1}{N} \sum_{\mathbf{k}} \left(\frac{\partial \varepsilon_{\mathbf{k}}}{\partial k_x} \right)^2 \delta(\varepsilon - \varepsilon_{\mathbf{k}})$ is the transport function, $A(\varepsilon, \nu) = -\frac{1}{\pi} \text{Im}[\nu + \mu - \varepsilon - \Sigma(\nu)]^{-1}$, and f is the Fermi function. For the square lattice $\sigma_0 = 2\pi$ and for triangular $\sigma_0 = 4\pi/\sqrt{3}$. For the calculation of conductivity in DMFT-QMC we need the real-frequency self-energy $\Sigma(\omega)$, which we obtain by Padé analytical continuation of the DMFT-QMC $\Sigma(i\omega_n)$. In the DMFT with NRG impurity solver (DMFT-NRG) we obtain the correlation functions directly on the real-frequency axis, but this method involves certain numerical approximations (see Appendix A).

In order to put into perspective the interaction strength $U = 10t$ and the temperature range that we consider, in Fig. 2 we sketch the paramagnetic (cluster) DMFT phase diagram for the triangular and square lattices at half-filling adapted from Refs. [46,47] (see also Refs. [36,37,48–54]). In the DMFT solution (blue lines) the critical interaction for the Mott metal-insulator transition (MIT) is $U_c \sim 2.5D$, where the half-bandwidth D is $4.5t$ and $4t$ for the triangular and the square lattice, respectively. The phase diagram features the region of coexistence of metallic and insulating solution below the critical end point at $T_c \approx 0.1t$. In this work we consider the temperatures above T_c . We set $U = 10t$, which is near U_c for the MIT in DMFT, but well within the Mott insulating part of the cluster DMFT and FTLM phase diagram.

III. RESULTS

We will first present the results for the thermodynamic properties in order to precisely identify the temperature range where the nonlocal correlations and finite-size effects are small or even negligible. In addition, from the thermodynamic quantities, e.g., from the specific heat, we can clearly identify the coherence temperature above which we observe the bad-metal transport regime. We then proceed with the key result

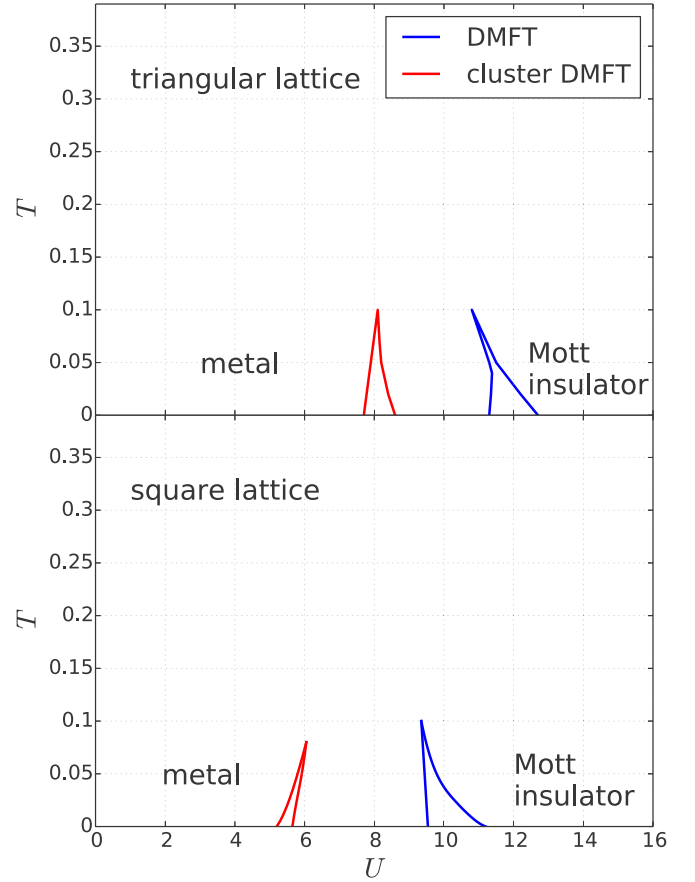


FIG. 2. Sketch of the paramagnetic phase diagram at half-filling, adapted from Refs. [46,47]. There is a region of the coexistence of metallic and insulating solution below the critical end point at T_c . The critical interaction is smaller in the cluster DMFT solution. Above T_c there is a gradual crossover from a metal to the Mott insulator. In this work we consider $T > T_c$ and $U = 10t$.

of this work by showing the contribution of vertex corrections to the resistivity and optical conductivity.

Before going into this detailed analysis, and in order to obtain a quick insight into the strength of nonlocal correlations, we compare in Fig. 3 the self-energy components in the cluster DMFT solution at two representative temperatures. We show the imaginary part of the DCA 4×4 self-energy at different patches of the Brillouin zone according to the color scheme of Fig. 1. The statistical error bar of the $\text{Im} \Sigma$ results presented in Fig. 3 we estimate by looking at the difference in $\text{Im} \Sigma$ between the last two iterations of the cluster DMFT loop. We monitor all \mathbf{K} points and the lowest three Matsubara frequencies. At lower temperature (bottom row), this difference is smaller than 0.05 (0.01) for the square (triangular) lattice, respectively. At higher temperature (upper row), these values are both 10 times lower and the error bar is much smaller than the size of the symbol. At $T = 0.4t$ the differences in the self-energy components are more pronounced on the square than on the triangular lattice, which goes along the general expectations that the larger connectivity ($z = 6$) and the frustrated magnetic fluctuations lead to the more local self-energy. At $T \sim 1.5t$ all the components of the self-energy almost coincide for both lattices. We note that for the triangular

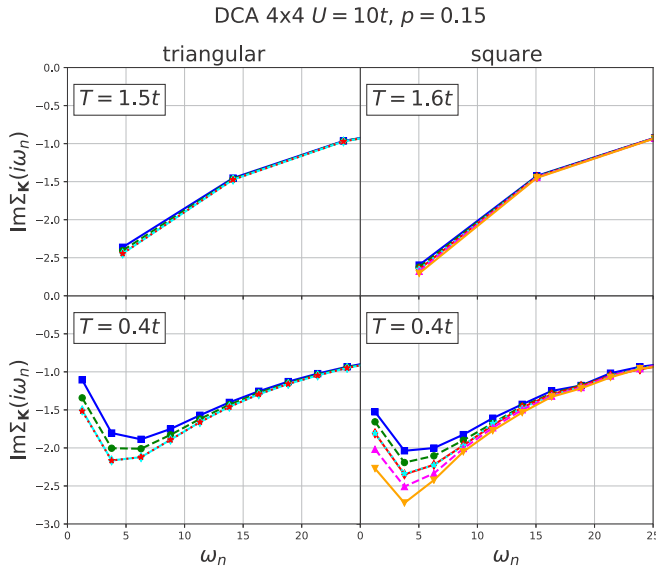


FIG. 3. Imaginary part of the self-energy at the Matsubara frequencies at different patches of the Brillouin zone for several temperatures for $p = 0.15$ hole doping. The position of the patches is indicated by the same colors as in Fig. 1. The solid lines are guide to the eye.

lattice the components of the self-energy marked by red and cyan colors are similar, but they do not coincide completely. There are four independent patches in this case. For the square lattice the red and cyan components of the self-energy are very similar, while we have six independent patches.

A. Thermodynamics

1. $p = 0.15$

We first show the results for hole doping $p = 0.15$. The results for the triangular lattice are shown in the left column of Fig. 4, and the results for the square lattice in the right column. Different rows correspond to the kinetic energy per lattice site E_{kin} , potential energy E_{pot} , total energy E_{tot} , specific heat $C = dE_{\text{tot}}/dT|_n$, and charge susceptibility χ_c . The DMFT results are shown with blue solid lines and FTLM with red dashed lines. The red circles correspond to DCA 4×4 , light green to DCA 2×2 , green to CDMFT 2×2 , and magenta to the CDMFT 2×1 result.

The FTLM results are shown down to $T = 0.2t$. The FTLM finite-size effects in thermodynamic quantities are small for $T \gtrsim 0.2t$ (see Appendix B). The DMFT results are shown for $T \gtrsim 0.05t$ and cluster DMFT for $T \gtrsim 0.2t$. Overall, the (cluster) DMFT and FTLM results for 15% doping look rather similar. The kinetic and potential energy do not differ much on the scale of the plots, and the specific heat looks similar.

The Fermi-liquid region, with $C \propto T$, is restricted to very low temperatures. For the triangular lattice we find a distinct maximum in $C(T)$ at $T \approx 0.4t$ in FTLM, and at $T \approx 0.3t$ in DMFT. This maximum is a signature of the coherence-incoherence crossover, when the quasiparticle peak in the density of states gradually diminishes and the bad-metal regime starts. The increase in the specific heat for $T \gtrsim 2t$ is

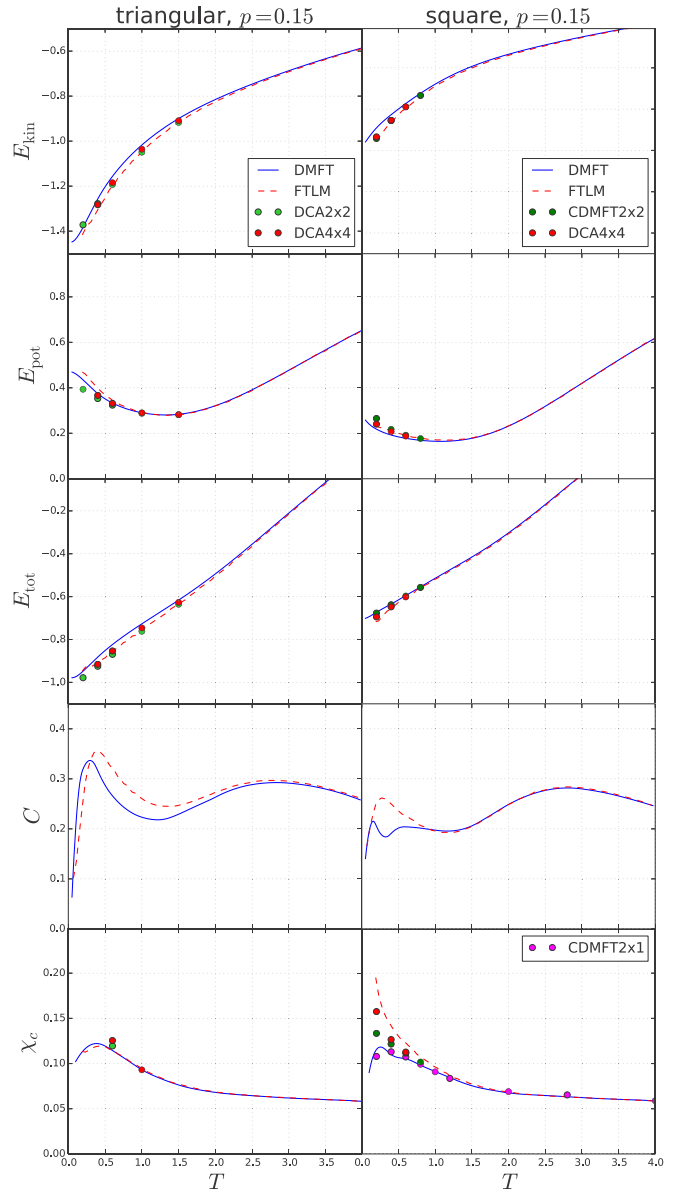


FIG. 4. Kinetic, potential, total energy, specific heat, and charge susceptibility as a function of temperature for the triangular and the square lattice at 15% doping.

caused by the charge excitations to the Hubbard band. The specific heat of the square lattice looks qualitatively the same. [A very small dip in the DMFT specific heat near $T = 0.4t$ for the square lattice may be an artifact of the numerics, where C is calculated by taking a derivative with respect to temperature of the interpolated $E_{\text{tot}}(T)$.] We note that the specific heat, shown here for the fixed particle density, is slightly different than the one for the fixed chemical potential $C_\mu = dE_{\text{tot}}/dT|_\mu$, as in Refs. [28,51,55].

For the square lattice all thermodynamic quantities obtained with different methods practically coincide for $T \gtrsim t$. This means that both the nonlocal correlations and the finite-size effects have negligible effect on thermodynamic quantities. For $T \lesssim t$ the DMFT and FTLM results start to differ. Interestingly, for the triangular lattice there is a small

difference in the DMFT and FTLM kinetic energy up to higher temperatures $T \sim 1.5t$. The FTLM and DCA 4×4 results coincide for $T \gtrsim t$, implying the absence of finite-size effects in the kinetic energy for both lattice types. We also note that the agreement of the CDMFT and DMFT solutions for the total energy on the square lattice at low temperatures is coincidental, as a result of a cancellation of differences in E_{kin} and E_{pot} .

The intersite correlations in the square lattice lead to an increase in the charge susceptibility at low temperatures (bottom panel in Fig. 4). Here, the FTLM and DCA 4×4 results are in rather good agreement. For the triangular lattice we found a sudden increase of χ_c at low temperatures in the DCA results (see Appendix B) but not in FTLM. These DCA points are not shown in Fig. 4 since we believe that they are an artifact of the particular choice of patching of the Brillouin zone. In order to keep the lattice symmetry, we had only four (in DCA 4×4) and two (in DCA 2×2) independent patches in the Brillouin zone for triangular lattice (Fig. 1). The average over twisted boundary conditions in FTLM reduces the finite-size error (see Appendix B), and hence we believe that the FTLM result for χ_c is correct down to $T = 0.2t$. We note that an increase of χ_c cannot be inferred from the ladder dual-fermion extension of DMFT [37] either. Still, further work would be needed to precisely resolve the low- T behavior of charge susceptibility for the triangular lattice.

2. $p = 0$

We now focus on thermodynamic quantities at half-filling (Fig. 5). In this case, the results can strongly depend on the method, especially since we have set the interaction to $U = 10t$, which is near the critical value for the Mott MIT in DMFT, while well within the insulating phase in the cluster DMFT and FTLM. The results with different methods almost coincide for $T \gtrsim 2t$ and are very similar down to $T \sim t$. The difference between the cluster DMFT and FTLM at half-filling is small, which means that the finite-size effects are small down to the lowest shown temperature $T = 0.2t$. Therefore, the substantial difference between the FTLM and single-site DMFT solutions at half-filling is mostly due to the absence of nonlocal correlations in DMFT.

The specific heat at half-filling is strongly affected by nonlocal correlations and lattice frustration. For triangular lattice the low-temperature maximum in $C(T)$ has different origin in the DMFT and FTLM solutions. The maximum in the FTLM is due to the low-energy spin excitations in frustrated triangular lattice, while in DMFT it is associated with the narrow quasiparticle peak since the DMFT solution becomes metallic as $T \rightarrow 0$. Our DMFT result agrees very well with the early work from Ref. [36] for $T \gtrsim t$. At lower temperatures there is some numerical discrepancy which we ascribe to the error due to the imaginary-time discretization in the Hirsch-Fye method used in that reference. For the square lattice the DMFT and FTLM solutions are both insulating. The maximum in the FTLM $C(T)$ is due to the spin excitations at energies $\sim 4t^2/U = 0.4t$, and it is absent in the paramagnetic DMFT solution which does not include dynamic nonlocal correlations. The increase in $C(T)$ at higher temperatures is due to the charge excitations to the upper Hubbard band.

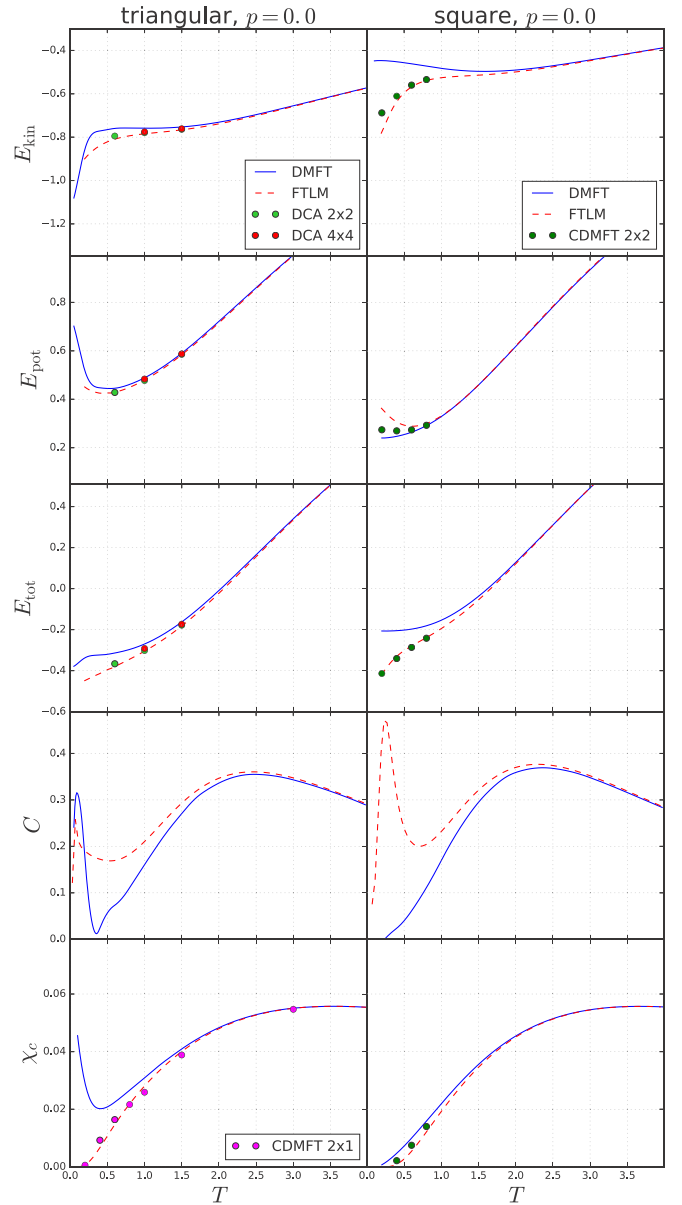


FIG. 5. Kinetic, potential, total energy, specific heat, and charge compressibility as a function of temperature for the triangular and the square lattice at half-filling.

B. Charge transport

The analysis of thermodynamic quantities has shown that the FTLM results for static quantities are close to exact down to $T \sim 0.5t$ or even $0.2t$. For charge transport we show the results for higher temperatures $T \gtrsim t$ since the finite-size effects are more pronounced in the current-current correlation function at lower temperatures.

An indication of the finite-size effects in optical conductivity can be obtained from the optical sum rule

$$\int_0^\infty d\omega \sigma(\omega) = \frac{\pi}{4V_{uc}} (-E_{\text{kin}}), \quad (6)$$

where V_{uc} is equal to 1 and $\frac{\sqrt{3}}{2}$ for the square and triangular lattice, respectively. The deviation from the sum rule in FTLM

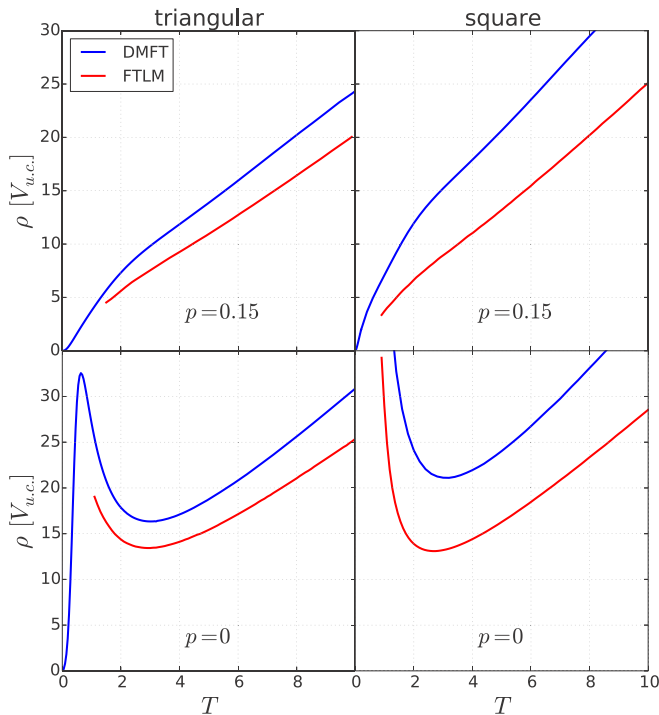
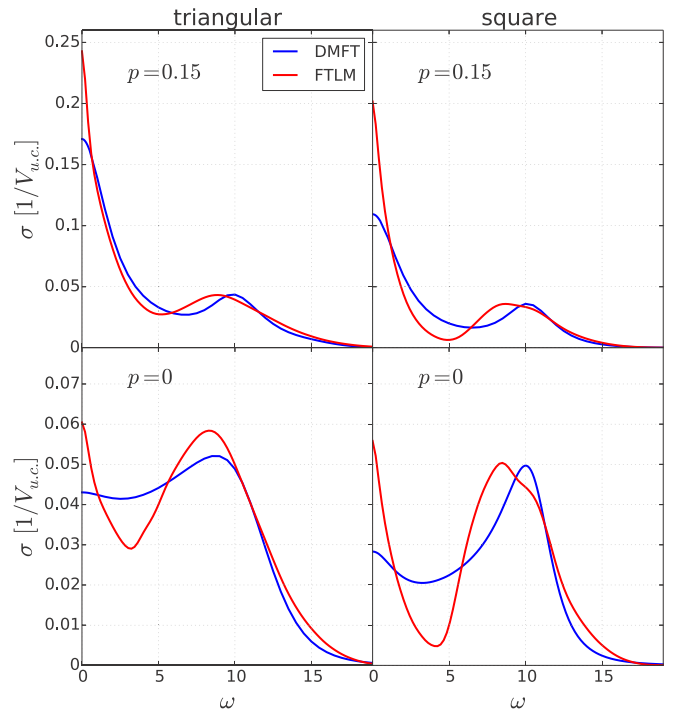


FIG. 6. Resistivity as a function of temperature.

can be ascribed to the finite charge stiffness and δ function at zero frequency in optical conductivity [28]. The FTLM result for dc resistivity, shown by the red lines in Fig. 6, corresponds the temperature range where the weight of the δ -function peak at zero frequency (charge stiffness) [28] is smaller than 0.5% of the total spectral weight. The other finite-size effects are small and the FTLM resistivity is expected to be close to the exact solution of the Hubbard model. The remaining uncertainty, due to the frequency broadening, is estimated to be below 10% (see Supplemental Material in Ref. [25]). Smallness of the finite-size effects for the square lattice at $T \gtrsim t$ was also confirmed from the current-current correlation function calculated on the 4×4 and 8×8 lattices using CTINT QMC (see Ref. [25]). For doped triangular lattice we show the conductivity data for $T \gtrsim 1.5t$ since below this temperature the weight of the charge stiffness δ function is larger than 0.5% of the total weight, which indicates larger finite-size effects.

The DMFT resistivity is shown in Fig. 6 by the blue lines. It is obtained using the NRG impurity solver. Numerical error of the DMFT-NRG method is small, as we confirmed by a comparison with the DMFT-QMC calculation followed by the Padé analytical continuation (see Appendix A). We note that we do not show the conductivity data in the DCA since in this approximation we cannot reliably calculate the conductivity beyond the bubble term. At high temperatures the bubble-term contribution in cluster DMFT does not differ from the one in single-site DMFT since the self-energy becomes local [25].

Since the FTLM resistivity in Fig. 6 is shown only for temperatures when both the nonlocal correlations and the finite-size effects are small, the difference between the DMFT and FTLM resistivity is due to the vertex corrections. Their contribution corresponds to the connected part of the current-current correlation function whereas the DMFT conductivity

FIG. 7. Optical conductivity at $T = 1.4$.

is given by the bubble diagram. A detailed analysis of vertex corrections for the square lattice is given in our previous work (Ref. [25]). Here, our main focus is on the comparison of the importance of vertex corrections for different lattices: the numerical results show that the vertex corrections to conductivity are less important in the case of the triangular lattice.

In the doped case, the FTLM solution gives the resistivity which is approximately linear in the entire temperature range shown in Fig. 6. This bad-metal linear- T temperature dependence is one of the key signatures of strong electronic correlations. The resistivity is here above the Mott-Ioffe-Regel limit which corresponds to the scattering length one lattice spacing within the Boltzmann theory. The Mott-Ioffe-Regel limit can be estimated as [6] $\rho_{\text{MIR}} \sim \sqrt{2\pi} \approx 2.5$.

At half-filling and low temperatures the result qualitatively depends on the applied method. For the half-filled triangular lattice at $U = 10t$ the DMFT solution gives a metal, whereas the nonlocal correlations lead to the Mott insulating state. Still, similar as for thermodynamic quantities, the numerically cheap DMFT gives an insulatinglike behavior and a rather good approximation down to $T \sim 0.5t$.

The optical conductivity, shown in Fig. 7 for $T = 1.4t$, provides further insight into the dependence of the vertex correction on the lattice geometry. The DMFT-QMC conductivity is calculated using Eq. (5) with $\Sigma(\omega)$ obtained by the Padé analytical continuation of $\Sigma(i\omega_n)$ (see Appendix A for a comparison with DMFT-NRG). In the DMFT solution, the Hubbard peak is determined by the single-particle processes and it is centered precisely at $\omega = U$. The vertex corrections in FTLM shift the position of the Hubbard peak to lower frequencies. The total spectral weight is the same in FTLM and DMFT solution since it obeys the sum rule of Eq. (6), while the kinetic energies coincide. The Ward identity for

vertex corrections [25,31]

$$\Lambda^{\text{conn}}(i\nu = 0) = -2T \frac{1}{N} \sum_{\mathbf{k}} v_{\mathbf{k}} \sum_{i\omega_n} G_{\mathbf{k}}^2(i\omega_n) \partial_{k_x} \Sigma_{\mathbf{k}}(i\omega_n) \quad (7)$$

also implies that the vertex corrections do not affect the sum rule if the self-energy is local. Here, $\Lambda(i\nu)$ is the current-current correlation function and $\Lambda(i\nu = 0) = \frac{1}{\pi} \int d\omega \sigma(\omega)$.

The results clearly show the much stronger effect of vertex corrections on the square lattice on all energy scales. In addition to a very different $\omega \rightarrow 0$ (dc) limit, we observe the more significant reduction of the Drude-like peak width and a larger shift of the Hubbard peak on the square lattice, with a more pronounced suppression of the optical weight at intermediate frequencies. We note that a broad low-frequency peak in conductivity is due to incoherent short-lived excitations characteristic of the bad-metal regime. The structure of the density of states in different transport regimes is discussed in Appendix C.

IV. CONCLUSION

In summary, we have performed a detailed comparison of the thermodynamic and charge transport properties of the Hubbard model on a triangular and square lattice. We identified the temperatures when the finite-size effects become negligible and the FTLM results on the 4×4 cluster are close to exact. In the doped case, for both lattice types, the resistivity is approximately linear in temperature for $T \gtrsim 1.5t$. In particular, we found that the contribution of vertex corrections to the optical and dc conductivity is smaller in the case of a triangular lattice, where it leads to $\sim 20\%$ decrease in dc resistivity as compared to the bubble term. The vertex corrections also leave a fingerprint on the position of the Hubbard peak in the optical conductivity, which is shifted from $\omega = U$ to slightly lower frequencies.

On general grounds, higher connectivity and/or magnetic frustration should lead to more local self-energy and smaller vertex corrections in the case of triangular lattice, as it is observed. However, the precise role of these physical mechanisms and possible other factors remains to be established. Another important open question is to find an efficient approximate scheme to evaluate the vertex corrections, which would be sufficiently numerically cheap to enable calculations of transport at lower temperatures and in real materials. These issues are to be addressed in the future, but we are now better positioned as we have established reliable results that can serve as a reference point.

With this work we also made a benchmark of several state-of-the-art numerical methods for solving the Hubbard model and calculating the conductivity at high temperatures. This may be a useful reference for calculations of conductivity using a recent approach that calculates perturbatively the correlation functions directly on the real-frequency axis [56–59], thus eliminating a need for analytical continuation, while going beyond the calculation on the 4×4 cluster.

ACKNOWLEDGMENTS

J.M. acknowledges useful discussions with F. Krien. A.V., J.V., and D.T. acknowledge funding provided by the Institute of Physics Belgrade, through the grant by

the Ministry of Education, Science, and Technological Development of the Republic of Serbia. J.K., R.Ž., and J.M. are supported by the Slovenian Research Agency (ARRS) under Programs No. P1-0044, No. J1-1696, and No. J1-2458. Numerical simulations were performed on the PARADOX supercomputing facility at the Scientific Computing Laboratory of the Institute of Physics Belgrade. The CTINT algorithm has been implemented using the TRIQS toolbox [60].

APPENDIX A: COMPARISON OF THE DMFT-NRG AND DMFT-QMC CONDUCTIVITY

Here, we compare the DMFT results for the dc resistivity and optical conductivity obtained with two different impurity solvers. The optical conductivity $\sigma(\omega)$ is calculated according to Eq. (5). The dc resistivity is equal to $\rho = \sigma^{-1}(\omega \rightarrow 0)$.

Within DMFT-NRG solver the self-energy is obtained directly on the real-frequency axis. There are three sources of errors in this approach: discretization errors, truncation errors, and (over)broadening errors. The method is based on the discretization of the continuum of states in the bath; the ensuing discretization errors can be reduced by performing the calculation for several different discretization meshes with interleaved points and averaging these results. It has been shown [45] that in the absence of interactions, the discretization error can be fully eliminated in a systematic manner. For an interacting problem, the cancellation of artifacts is only approximate, but typically very good, so that this is a minor source of errors. The truncation errors arise because in the iterative diagonalization one discards high-energy states after each set of diagonalizations. For static quantities this error is negligible, but it affects the dynamical (frequency-resolved) quantities because they are calculated from contributions linking kept and discarded states [61–63]. Finally, the raw spectral function in the form of δ peaks needs to be broadened in order to obtain the smooth spectrum. If the results are overbroadened, this can result in a severe overestimation of resistivity, and this is typically the main source of error in the NRG for this quantity. Fortunately, the resistivity is calculated as an integrated quantity, thus, the broadening kernel width can be systematically reduced [20,64]. The lower limit is set by the possible convergence issues in the DMFT self-consistency cycle due to jagged aspect of all quantities, where the actual limit value is problem dependent. In the NRG results reported in this work, it was possible to use very narrow broadening kernel. By studying the dependence of the $\rho(T)$ curves on the kernel width, we estimate that the presented results have at most a few percent error even at the highest temperatures considered.

The DMFT-QMC gives the self-energy $\Sigma(i\omega_n)$ at the Matsubara frequencies and the analytical continuation is necessary to obtain $\Sigma(\omega)$. The statistical error in QMC makes the analytical continuation particularly challenging. However, at high temperatures the CTINT QMC algorithm is very efficient. Running a single DMFT iteration for 10 minutes on 128 cores and using 20 or more iterations, we obtained the self-energies with the statistical error $|\delta \Sigma(i\omega_0)| \approx 5 \times 10^{-4}$ and $|\delta G(i\omega_0)| \approx 2 \times 10^{-5}$ at the first Matsubara frequency at $T = t$. Such a small statistical error makes the Padé analytical continuation possible for temperatures $T \lesssim 2t$.

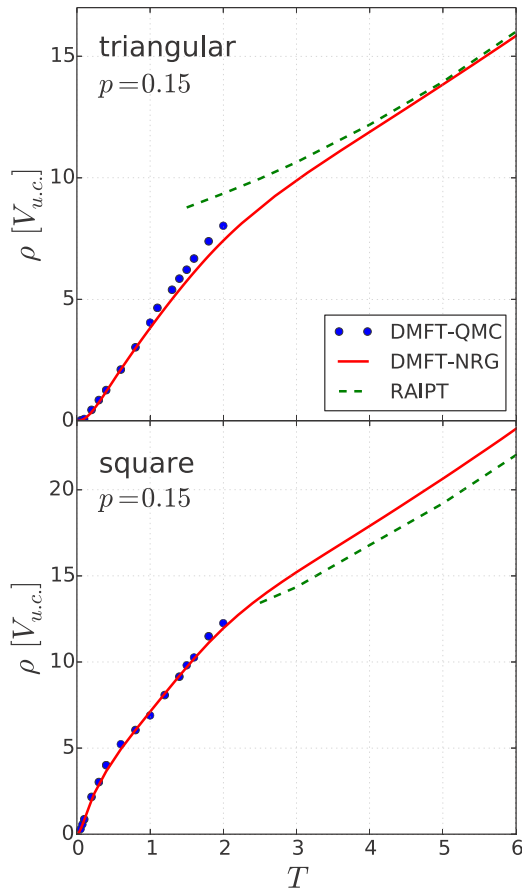


FIG. 8. DMFT-QMC (blue dots) and DMFT-NRG (red lines) resistivity as a function of temperature. The analytical continuation of the self-energy is performed with the Padé method. At high temperatures the DMFT-NRG result agrees rather well with the RAIPT (green dashed lines).

We have checked that Padé continuation gives similar results for $\Sigma(\omega)$ when performed on $\Sigma(i\omega_n)$ taken from last few DMFT iterations. We then used $\Sigma(i\omega_n)$ averaged over the last five iterations to further reduce the noise in $\Sigma(i\omega_n)$, before performing the Padé analytical continuation subsequently used in the calculation of the conductivity. We also obtained $G(\omega)$ directly by the Padé analytical continuation of $G(i\omega_n)$, and checked that the result is consistent with the one calculated as $G(\omega) = \int d\varepsilon \rho_0(\varepsilon)[\omega + \mu - \varepsilon - \Sigma(\omega)]^{-1}$. These cross checks have confirmed that Padé analytical continuation is rather reliable.

Figure 8 shows the temperature dependence of resistivity calculated with the DMFT-NRG (red lines) and DMFT-QMC (blue dots). For the square lattice we find excellent agreement between the two methods. For the triangular lattice we find some discrepancy for $T \sim 1.5t$, which is likely due to the approximations in DMFT-NRG. We also find that the real-axis iterative perturbation theory [65–67] (RAIPT) agrees rather well with the DMFT-NRG solution for $T \gtrsim 2t$.

It is also interesting to note how the lattice geometry can influence the range of the Fermi liquid $\rho \propto T^2$ behavior in the DMFT solution. In the DMFT equations the lattice structure enters only through the noninteracting density of states. We

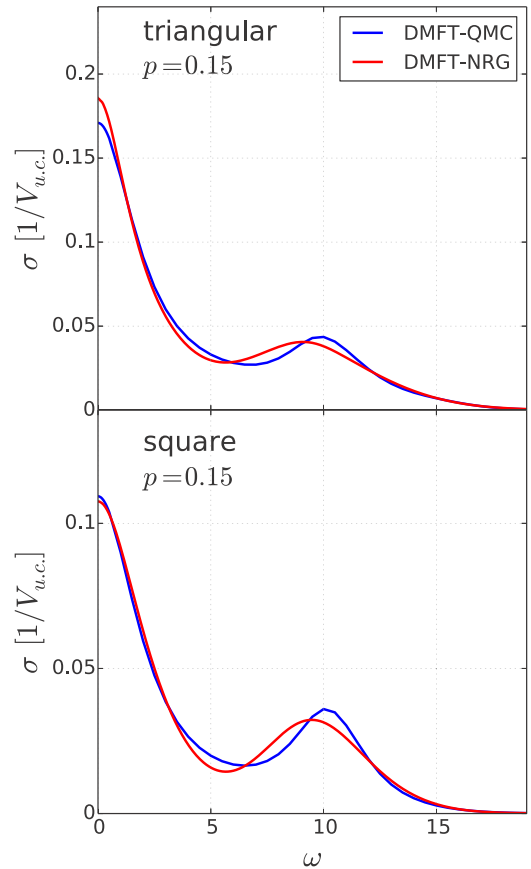


FIG. 9. DMFT-QMC and DMFT-NRG optical conductivity at $T = 1.4t$.

observe $\rho \propto T^2$ behavior up to much lower temperatures on the square lattice. In this case, $\rho \propto T^2$ region is hardly visible on the scale of the plot, while $\rho \propto T^2$ up to $T \sim 0.3t$ on the triangular lattice. This observation is in agreement with the extension of the $C \propto T$ region in $C(T)$, which is restricted to lower temperatures in the case of a square lattice (Fig. 4).

A comparison of the DMFT-NRG (red lines) and DMFT-QMC (blue lines) optical conductivity at $T = 1.4t$ is shown in Fig. 9. The overall agreement is very good. We, however, find a small discrepancy at $\omega \sim 10t$. The DMFT-QMC result has the Hubbard peak in $\sigma(\omega)$ centered exactly at $\omega = U$, whereas it is shifted to slightly lower frequency in the DMFT-NRG solution. This shift is an artifact of numerical approximations in DMFT-NRG. A position of the Hubbard peak at $U = 10t$ is another manifestation of the precision of analytical continuation of the QMC data.

APPENDIX B: FINITE-SIZE EFFECTS IN CHARGE SUSCEPTIBILITY

In Fig. 10 we show the charge susceptibility obtained with different methods. The single-site DMFT result agrees very well with the 4×4 FTLM after averaging over the twisted boundary conditions. We show χ_c averaged over $N_{\text{tbc}} = 1, 4, 16, 64,$ and 128 clusters with different boundary conditions. χ_c obtained with a single setup of boundary conditions deviates at low temperatures from the averaged values. The DCA results for $T \lesssim 0.5t$ are also inconsistent.

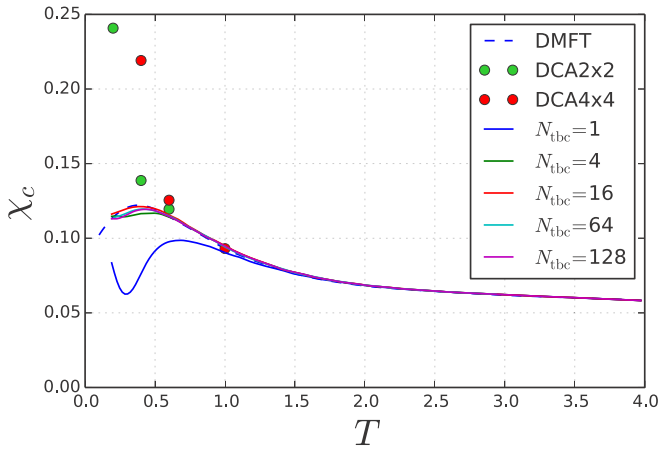


FIG. 10. Charge susceptibility as a function of temperature for the triangular lattice at $p = 0.15$ hole doping.

We believe that this is an artifact of the particular choice of the Brillouin zone patches. In DCA 4×4 and 2×2 we have just four and two independent patches in the Brillouin zone for triangular lattice, respectively.

APPENDIX C: DMFT DENSITY OF STATES

Here, we illustrate the density of states in different transport regimes in the DMFT solution. The results in Fig. 11 are obtained with the QMC solver followed by the Padé analytical continuation. We have checked that the density of states agrees with the DMFT-NRG result.

In the Fermi-liquid regime at low temperatures there is a peak in the density of states around the Fermi level. In the doped case the coherence-decoherence crossover is at temperature $T \sim 0.3$, as we established from the specific-heat data (see Fig. 4) and from the condition that the resistivity reaches the Mott-Ioffe-Regel limit (see Sec. III B). In agreement with earlier work [10,12], we see that at $T \sim 0.3$ there is a peak in the density of states even though long-lived quasiparticles are absent. At even higher temperatures (here shown $T = 1.4$), deeply in the bad-metal regime, the peak at the density of states at the Fermi level is completely washed out.

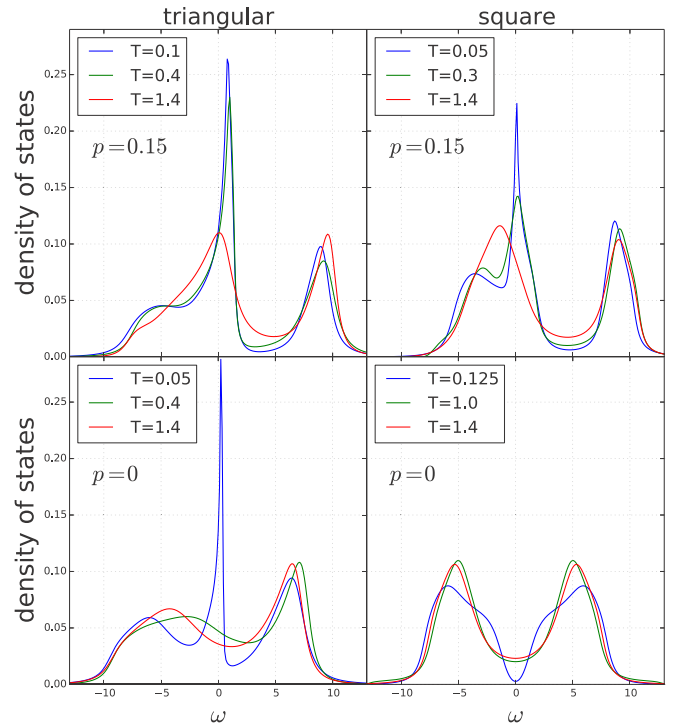


FIG. 11. Density of states in the Fermi liquid at low temperatures and in the bad-metal regime at high temperatures.

At half-filling the result is very sensitive to the exact position of parameters on the U - T phase diagram (see Fig. 2). For the triangular lattice at $U = 10$ the solution is metallic even at low temperature which leads to the formation of narrow quasiparticle peak at the Fermi level. This peak is quickly suppressed by thermal fluctuations which is accompanied by a sudden increase in the resistivity. For the square lattice at $U = 10$ the system is insulating above for $T \gtrsim 0.03$, while the Mott gap gradually gets filled as the temperature increases. We note that the low-temperature peak in optical conductivity in Fig. 7 is not connected to the existence of quasiparticles. It is just a consequence of a finite spectral density at the Fermi level (the absence of an energy gap), as expected in the bad-metal regime.

-
- [1] S. A. Kivelson, I. P. Bindloss, E. Fradkin, V. Oganesyan, J. M. Tranquada, A. Kapitulnik, and C. Howald, *Rev. Mod. Phys.* **75**, 1201 (2003).
 - [2] B. J. Powell and R. H. McKenzie, *Rep. Prog. Phys.* **74**, 056501 (2011).
 - [3] K. Miyagawa, A. Kawamoto, Y. Nakazawa, and K. Kanoda, *Phys. Rev. Lett.* **75**, 1174 (1995).
 - [4] Y. Shimizu, K. Miyagawa, K. Kanoda, M. Maesato, and G. Saito, *Phys. Rev. Lett.* **91**, 107001 (2003).
 - [5] V. Dobrosavljević, N. Trivedi, and J. M. Valles, Jr., *Conductor-Insulator Quantum Phase Transitions* (Oxford University Press, Oxford, 2012).
 - [6] O. Gunnarsson, M. Calandra, and J. E. Han, *Rev. Mod. Phys.* **75**, 1085 (2003).
 - [7] N. E. Hussey, K. Takenaka, and H. Takagi, *Philos. Mag.* **84**, 2847 (2004).
 - [8] M. M. Qazilbash, K. S. Burch, D. Whisler, D. Shrekenhamer, B. G. Chae, H. T. Kim, and D. N. Basov, *Phys. Rev. B* **74**, 205118 (2006).
 - [9] M. M. Qazilbash, J. J. Hamlin, R. E. Baumbach, L. Zhang, D. J. Singh, M. B. Maple, and D. N. Basov, *Nat. Phys.* **5**, 647 (2009).
 - [10] X. Deng, J. Mravlje, R. Žitko, M. Ferrero, G. Kotliar, and A. Georges, *Phys. Rev. Lett.* **110**, 086401 (2013).
 - [11] W. Xu, K. Haule, and G. Kotliar, *Phys. Rev. Lett.* **111**, 036401 (2013).
 - [12] J. Vučković, D. Tanasković, M. J. Rozenberg, and V. Dobrosavljević, *Phys. Rev. Lett.* **114**, 246402 (2015).

- [13] H. Terletska, J. Vučićević, D. Tanasković, and V. Dobrosavljević, *Phys. Rev. Lett.* **107**, 026401 (2011).
- [14] J. Vučićević, H. Terletska, D. Tanasković, and V. Dobrosavljević, *Phys. Rev. B* **88**, 075143 (2013).
- [15] T. Furukawa, K. Miyagawa, H. Taniguchi, R. Kato, and K. Kanoda, *Nat. Phys.* **11**, 221 (2015).
- [16] H. Eisenlohr, S.-S. B. Lee, and M. Vojta, *Phys. Rev. B* **100**, 155152 (2019).
- [17] B. H. Moon, G. H. Han, M. M. Radonjić, H. Ji, and V. Dobrosavljević, [arXiv:1911.02772](https://arxiv.org/abs/1911.02772).
- [18] J. Kokalj, *Phys. Rev. B* **95**, 041110(R) (2017).
- [19] E. W. Huang, R. Sheppard, B. Moritz, and T. P. Devereaux, *Science* **366**, 987 (2019).
- [20] E. Perepelitsky, A. Galatas, J. Mravlje, R. Žitko, E. Khatami, B. S. Shastry, and A. Georges, *Phys. Rev. B* **94**, 235115 (2016).
- [21] S. Hartnoll, *Nat. Phys.* **11**, 54 (2015).
- [22] S. A. Hartnoll, A. Lucas, and S. Sachdev, *Holographic Quantum Matter* (MIT Press, Cambridge, MA, 2018).
- [23] P. Cha, A. A. Patel, E. Gull, and E.-A. Kim, [arXiv:1910.07530](https://arxiv.org/abs/1910.07530).
- [24] P. T. Brown, D. Mitra, E. Guardado-Sanchez, R. Nourafkan, A. Reymbaut, C.-D. Hébert, S. Bergeron, A.-M. S. Tremblay, J. Kokalj, D. A. Huse, P. Schauß, and W. S. Bakr, *Science* **363**, 379 (2019).
- [25] J. Vučićević, J. Kokalj, R. Žitko, N. Wentzell, D. Tanasković, and J. Mravlje, *Phys. Rev. Lett.* **123**, 036601 (2019).
- [26] A. Georges, G. Kotliar, W. Krauth, and M. J. Rozenberg, *Rev. Mod. Phys.* **68**, 13 (1996).
- [27] T. A. Maier, M. Jarrell, T. Pruschke, and M. H. Hettler, *Rev. Mod. Phys.* **77**, 1027 (2005).
- [28] J. Jaklič and P. Prelovšek, *Adv. Phys.* **49**, 1 (2000).
- [29] N. Lin, E. Gull, and A. J. Millis, *Phys. Rev. B* **80**, 161105(R) (2009).
- [30] N. Lin, E. Gull, and A. J. Millis, *Phys. Rev. B* **82**, 045104 (2010).
- [31] D. Bergeron, V. Hankevych, B. Kyung, and A.-M. S. Tremblay, *Phys. Rev. B* **84**, 085128 (2011).
- [32] T. Sato, K. Hattori, and H. Tsunetsugu, *Phys. Rev. B* **86**, 235137 (2012).
- [33] T. Sato and H. Tsunetsugu, *Phys. Rev. B* **94**, 085110 (2016).
- [34] A. Kauch, P. Pudleiner, K. Astleithner, P. Thunström, T. Ribic, and K. Held, *Phys. Rev. Lett.* **124**, 047401 (2020).
- [35] A. Georges, *Ann. Phys. (Berlin)* **523**, 672 (2011).
- [36] K. Aryanpour, W. E. Pickett, and R. T. Scalettar, *Phys. Rev. B* **74**, 085117 (2006).
- [37] G. Li, A. E. Antipov, A. N. Rubtsov, S. Kirchner, and W. Hanke, *Phys. Rev. B* **89**, 161118(R) (2014).
- [38] G. Kotliar, S. Y. Savrasov, G. Pálsson, and G. Biroli, *Phys. Rev. Lett.* **87**, 186401 (2001).
- [39] G. Biroli and G. Kotliar, *Phys. Rev. B* **65**, 155112 (2002).
- [40] A. N. Rubtsov and A. I. Lichtenstein, *J. Exp. Theor. Phys. Lett.* **80**, 61 (2004).
- [41] E. Gull, A. J. Millis, A. I. Lichtenstein, A. N. Rubtsov, M. Troyer, and P. Werner, *Rev. Mod. Phys.* **83**, 349 (2011).
- [42] K. G. Wilson, *Rev. Mod. Phys.* **47**, 773 (1975).
- [43] H. R. Krishna-murthy, J. W. Wilkins, and K. G. Wilson, *Phys. Rev. B* **21**, 1003 (1980).
- [44] R. Bulla, T. A. Costi, and T. Pruschke, *Rev. Mod. Phys.* **80**, 395 (2008).
- [45] R. Žitko and T. Pruschke, *Phys. Rev. B* **79**, 085106 (2009).
- [46] H. T. Dang, X. Y. Xu, K.-S. Chen, Z. Y. Meng, and S. Wessel, *Phys. Rev. B* **91**, 155101 (2015).
- [47] H. Park, K. Haule, and G. Kotliar, *Phys. Rev. Lett.* **101**, 186403 (2008).
- [48] H. Lee, G. Li, and H. Monien, *Phys. Rev. B* **78**, 205117 (2008).
- [49] T. Shirakawa, T. Tohyama, J. Kokalj, S. Sota, and S. Yunoki, *Phys. Rev. B* **96**, 205130 (2017).
- [50] J. Merino, B. J. Powell, and R. H. McKenzie, *Phys. Rev. B* **73**, 235107 (2006).
- [51] J. Kokalj and R. H. McKenzie, *Phys. Rev. Lett.* **110**, 206402 (2013).
- [52] T. Schäfer, F. Geles, D. Rost, G. Rohringer, E. Arrigoni, K. Held, N. Blümer, M. Aichhorn, and A. Toschi, *Phys. Rev. B* **91**, 125109 (2015).
- [53] E. G. C. P. van Loon, M. I. Katsnelson, and H. Hafermann, *Phys. Rev. B* **98**, 155117 (2018).
- [54] C. Walsh, P. Sémon, D. Poulin, G. Sordi, and A.-M. S. Tremblay, *Phys. Rev. B* **99**, 075122 (2019).
- [55] J. Bonča and P. Prelovšek, *Phys. Rev. B* **67**, 085103 (2003).
- [56] J. Vučićević and M. Ferrero, *Phys. Rev. B* **101**, 075113 (2020).
- [57] A. Taheridehkordi, S. H. Curnoe, and J. P. F. LeBlanc, *Phys. Rev. B* **99**, 035120 (2019).
- [58] A. Taheridehkordi, S. H. Curnoe, and J. P. F. LeBlanc, *Phys. Rev. B* **101**, 125109 (2020).
- [59] A. Taheridehkordi, S. H. Curnoe, and J. P. F. LeBlanc, *Phys. Rev. B* **102**, 045115 (2020).
- [60] O. Parcollet, M. Ferrero, T. Ayrat, H. Hafermann, P. Seth, and I. S. Krivenko, *Comput. Phys. Commun.* **196**, 398 (2015).
- [61] R. Peters, T. Pruschke, and F. B. Anders, *Phys. Rev. B* **74**, 245114 (2006).
- [62] A. Weichselbaum and J. von Delft, *Phys. Rev. Lett.* **99**, 076402 (2007).
- [63] R. Žitko, *Phys. Rev. B* **84**, 085142 (2011).
- [64] R. Žitko, D. Hansen, E. Perepelitsky, J. Mravlje, A. Georges, and B. S. Shastry, *Phys. Rev. B* **88**, 235132 (2013).
- [65] H. Kajueter and G. Kotliar, *Phys. Rev. Lett.* **77**, 131 (1996).
- [66] M. Potthoff, T. Wegner, and W. Nolting, *Phys. Rev. B* **55**, 16132 (1997).
- [67] L.-F. Arsenault, P. Sémon, and A.-M. S. Tremblay, *Phys. Rev. B* **86**, 085133 (2012).

UNIVERSITY OF BELGRADE
FACULTY OF PHYSICS

Ana Vranić

**EVOLVING COMPLEX NETWORKS:
STRUCTURE AND DYNAMICS**

Doctoral Dissertation

Belgrade, 2023

УНИВЕРЗИТЕТ У БЕОГРАДУ
ФИЗИЧКИ ФАКУЛТЕТ

Ана Вранић

**РАСТУЋЕ КОМПЛЕКСНЕ МРЕЖЕ:
СТРУКТУРА И ДИНАМИКА**

докторска дисертација

Београд, 2023

Thesis Defense Committee

Thesis advisor:

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

Committee members:

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

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

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

Acknowledgements

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

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

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

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

Abstract

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

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

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

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

The complex network theory allows us to determine how different network properties evolve and understand how this evolution influences their sustainability. We use data from Stack Exchange sites and compare the evolution of network structure for pairs of active and closed communities during their early phase of existence. Stack Exchange sites are question-and-answer platforms where users share knowledge on some specific topic. We compare active and closed communities on four topics, namely astronomy, literature, economics, and physics. We analyze the structural patterns in these communities and find that active ones are more clustered and characterized by better-connected and stable cores. Core users are crucial for a healthy community and need to be trustworthy. Through the dynamic reputation model, we measure the level of trust in these communities. In active communities, core users show a higher reputation than in closed communities, indicating the importance that a stable core develops early and has a high level of trust.

Keywords: statistical physics of complex systems, the structure and dynamics of complex networks, modeling online social systems

Research field: Physics

Research subfield: Statistical physics

UDC number: 536

Сажетак

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

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

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

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

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

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

Кључне речи: статистичка физика комплексних система, структура и динамика комплексних мрежа, моделовање онлајн социјалних система

Научна област: Физика

Ужа научна област: Статистичка физика

УДК број: 536

Contents

Acknowledgements	iii
Abstract	v
Contents	ix
List of figures	xi
List of Tables	xii
1 Introduction	1
1.1 Complex networks	4
1.2 Thesis outline	7
2 Methodology	9
2.1 The measures of complex network structure	9
2.2 Community structure	12
2.3 The probability distributions	17
2.4 Network models	21
2.5 Fractal analysis	28
2.6 Dynamical reputation model	32
3 Evolving complex network structure dependence on the properties of growth signals	35
3.1 Aging network model with growth signal	35
3.2 Long range correlated signals	41
3.3 Conclusions	43
4 The growth of social groups	45
4.1 Empirical analysis of the social group growth	45
4.2 Theoretical model of social group growth	49
4.3 The growth of real social groups	53
4.4 Conclusions	58
5 The sustainability of evolving knowledge-based communities	59
5.1 Network properties of Stack Exchange data	60
5.2 Core-periphery structure	62
5.3 Dynamical Reputation on Stack Exchange communities	65

5.4	Conclusions	68
6	Conclusions	71
A	Stack Exchange	75
A.1	Comparison between active and closed SE communities	76
B	Selection of Dynamical Reputation Model parameters	79
C	The choice of the sliding window	83
D	Robustness of core-periphery algorithm	85
	Bibliography	89
	Biography of the author	99

List of figures

1.1	Konigsberg problem of seven bridges.	2
1.2	Graph, matrix and edge list representations.	4
1.3	Different network representations.	5
1.4	Bipartite network.	6
1.5	Temporal network.	7
2.1	Different communities structures.	13
2.2	Probability distributions on a linear and double logarithmic scale.	18
2.3	Erdős-Rényi graph.	22
2.4	Degree distribution of Erdős-Rényi graph.	23
2.5	Watts and Strogatz graph model creation.	24
2.6	Barabasi-Albert model.	25
2.7	Aging model.	27
2.8	Phase diagram of aging network model.	27
2.9	Multifractal, monofractal and white noise signals.	30
2.10	Detrending multifractal signal.	30
2.11	Fluctuating function and Hurst exponent.	31
2.12	User reputations.	33
3.1	Nonlinear growth of the network.	36
3.2	Properties of MySpace signal.	37
3.3	Properties of the TECH and Poisson signals.	37
3.4	D-measure for networks generated with real signals.	39
3.5	Structural properties of networks.	40
3.6	Long range correlated monofractal signals.	42
3.7	D-distance for networks generated with monofractal signals.	42
3.8	Assortativity index and mean clustering coefficient.	43
4.1	Properties of Meetup and Subreddit groups.	47
4.2	Universality in the Meetup and Reddit groups.	48
4.3	Bipartite groups growth model.	50
4.4	Group size distribution for different model parameters.	51
4.5	Comparison between preferential and random linking in the groups' growth model.	52
4.6	The estimation of the model parameters for a groups growth model.	53
4.7	The comparison between empirical and simulated data.	55
4.8	The fitting of empirical group size distributions.	56
4.9	The fitting of simulated group size distributions.	57

4.10	Users degree distribution	57
5.1	Degree distribution of Stack Exchange websites.	60
5.2	Neighbor degree dependence on the node degree of Stack Exchange websites.	61
5.3	Clustering coefficient dependence on the node degree of Stack Exchange websites.	61
5.4	Mean clustering coefficient of Stack Exchange websites.	62
5.5	Number of links per node of Stack Exchange websites.	63
5.6	The size of the core of Stack Exchange websites.	63
5.7	Jaccard index between core users of Stack Exchange websites.	64
5.8	Mean Jaccard index between core users of Stack Exchange websites.	64
5.9	Number of links per node of Stack Exchange websites.	65
5.10	Number of active users and dynamic reputation of Stack Exchange websites.	65
5.11	Dynamical reputation within core of Stack Exchange websites.	66
5.12	Ratio between the total reputation within network core and periphery of Stack Exchange websites.	67
5.13	Gini index of dynamic reputation of Stack Exchange websites.	67
5.14	Dynamic reputation assortativity of Stack Exchange websites.	67
5.15	Coefficient of correlation between users' dynamic reputation of Stack Exchange websites.	68
A.1	Number of active questions within seven days sliding windows.	76
B.1	Single users reputations.	79
B.2	RMSE between the number of users in 30 days sliding window and positive reputation.	80
B.3	Number of users in 30 days sliding window and positive reputation.	81
B.4	Number of users in Stack Exchange community who remain to be active.	81
C.1	Stack Exchange properties for different sliding window.	84
D.1	Stability of the core-periphery structures.	87

List of Tables

4.1	Jensen Shannon divergence between group sizes distributions from model and data.	54
4.2	The likelihood ratio R and p-value for fitting empirical data.	56
4.3	The likelihood ratio R and p-value for fitting simulated data.	56
A.1	Percentage of negatively voted interactions.	75
A.2	Community overview for first 180 days.	76
A.3	Community overview for first 180 days according to SE criteria.	77

Chapter 1

Introduction

Many real systems, such as brain networks, social organizations, cities, or cells, consist of many interacting units and belong to a class commonly known as complex systems. One of the most prominent characteristics of complex systems is that they exhibit emergent collective behavior that can not be predicted based on the behavior of individual components. The interactions between system components can be represented as a complex network [1]. The emergence of collective behavior strongly depends on the structure of the network of interactions. The structure of the brain network and its properties are fundamental for brain functioning, while an emergent phenomenon is human intelligence. In societies, people's interactions lead to civilization, economy, and formation of social groups [2]. Also, the animal populations show different levels of organization: such as patterns in bird flocks or schools of fish [2].

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

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

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

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

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

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

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

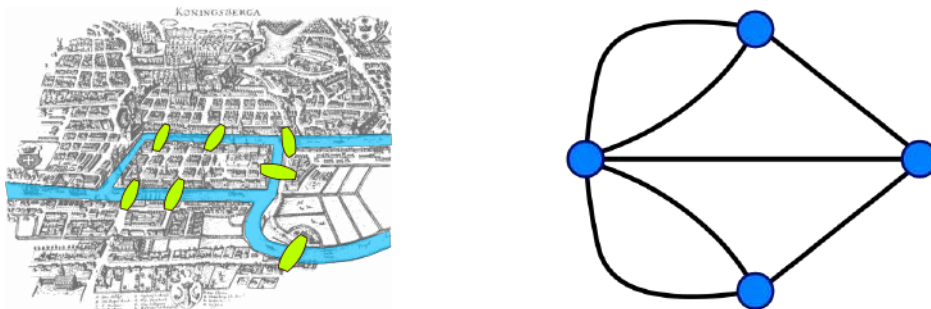


Figure 1.1: The Konigsberg problem of seven bridges. The left panel shows the original map of the bridges; the right panel shows its graph representation.

Until the late 1990s, graph theory was not widely used. Back then, the most crucial model was the Erdos-Reni model of random graphs, which considers a fixed number of nodes in the network connected randomly, resulting in the Poisson degree distribution. When researchers got an idea to map the World Wide Web (WWW) on the network and analyze its properties [27], they found that degree distribution follows the power-law contrary to expected behavior from random graph model

[28]. Because the power-law distribution is the same on all scales, such networks are called scale-free. Besides the scale-free property, empirical analysis of various complex networks showed the small-world property and the high clustering coefficient [29, 30]. Two seminal papers from 1999 inspired further research in complex networks. Watts and Strogatz [31] proposed the model where rewiring of edges on regular lattice leads to the network in which paths between any two nodes become short (small-world) and nodes become densely connected, resulting in a high clustering coefficient. On the other hand, Barabasi and Albert (BA) [32] introduced the model, where the network grows over time, and the new nodes tend to connect high-degree nodes; it produces scale-free networks with few highly connected nodes.

Different complex network models were proposed to describe the structure and dynamics of social and technological systems. The node degree is one of many node features that determine the linking probability, and the linking probability may be nonlinear in node degree or may depend on the age of the node [33, 34]. In the BA model, the links are introduced through new nodes, so it was proposed that links can be created between existing nodes in the network.

Furthermore, the BA model considers the constant network growth, where a fixed number of nodes is added at each step. The research on various social systems shows time-dependent growth, and we record the exponential growth of online systems [35]. Some models considered that nodes become inactive or even that network grows through a nonlinear number of links [36]. On the other hand, models with accelerated growth in the number of nodes [37] simulate exponential expansion of the online social systems. But the growth is not only accelerated; the time series of new nodes has trends and reflect the typical human behavior [38, 39, 40].

Research has also been devoted to using generated networks to analyze dynamic processes on top of them. Central questions are about the spread of epidemics, information diffusion, or emotional interactions among elements [18]. These systems are modeled using agent-based models, while the robustness is often studied by percolation and diffusion phenomena in complex networks. It was shown that scale-free networks' connectivity is sensitive to removing highly connected nodes. On the other hand, eliminating small degree nodes won't affect the scale-free structure [41]. They also show resilience to random attacks. Real-world networks are often characterized by community structure. They are common for social networks, where people with similar interests group together. Mostly adopted definition of a community is a group of densely connected nodes. The complex network theory provides different models for generating networks with community structure but also develops the algorithms for inferring the community structure from the underlying network.

The complex network models contribute to our knowledge, connecting the network topology and the dynamics of the system and helping us to understand underlying mechanisms that lead to the emergence of the properties of the complex networks [32, 42, 43, 44]. Complex network models must gain insights based on empirical data and social theories, and they are data-driven and require the development of computational approaches. The physicists showed interest in modeling complex systems by applying statistical physics approaches. Recently, the theory of graph neural networks (GNN) emerged from computer science, where machine learning methods are found helpful in inferring the properties of the network [45, 46, 47]. For example, they are used to determine missing links and recommend to users in online social networks [48, 49] or to develop generative GNN models that lead to the discovery of new drugs [50, 51].

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

1.1 Complex networks

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

Even though graphical representation can be useful for describing the network structure, numerical representation allows us to characterize the statistical properties of the networks. The graph G , with N nodes could be represented with **adjacency matrix** $|A| = N \times N$ [12]. The matrix elements are equal to 1 if there is a connection between two nodes v_i and v_j :

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

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

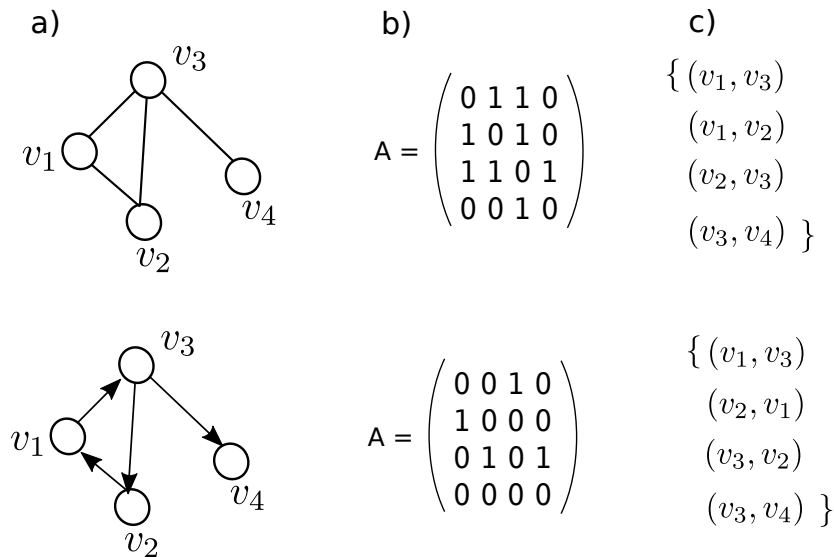


Figure 1.2: a) Graph representation of undirected (top panel) and directed (bottom panel) network. The same networks are represented with adjacency matrices in column b) and edge list representation in column c).

The number of edges and nodes are dependent variables. Considering that each node can make $N - 1$ connections, the maximum number of the edges in the network is $L_{max} = N(N - 1)/2$, as each edge is counted twice. For a directed network, it is possible to draw $L_{max} = N(N - 1)$

edges [52]. When it comes to large networks, they are sparse, meaning that the number of links is $L \ll L_{max}$. Consequently, the adjacency matrix is also a sparse structure (has many zeros) that takes a large portion of computer memory [53]. It is common to represent the graph as an edge list. In this case, illustrated in Figure 1.2, column c), a graph is described with the list of links that are in the graph, $G = \{\{v_i, v_j\}\}$. Still, with this representation, we cannot distinguish between directed and undirected graph structures, so the computational algorithm should specify if the edges are symmetric or not.

Sometimes is essential to include the specific properties of the system in the network representation. For example, to emphasize the frequent interactions between nodes, edges can be assigned with different values; such networks are **weighted**. In a collaboration network, authors who collaborate more often have stronger interaction. They can be described with an adjacency matrix, whose elements can take any real number $A_{ij} = w_{ij}$ and $w_{ij} > 0$. In general, edges may be associated with any categorical variable. Similarly, properties can be added to nodes or the whole network structure. Edges could be characterized by the time when the interaction between nodes happens, which includes the **temporal** component in the network representation, as in phone calls networks. Finally, if two nodes interact differently, the **multigraph** is an appropriate configuration where multiple edges are allowed. The transportation network, consisting of roads and railways, could be seen as a multigraph. Figure 1.3 presents the graphical representation of discussed network representations.

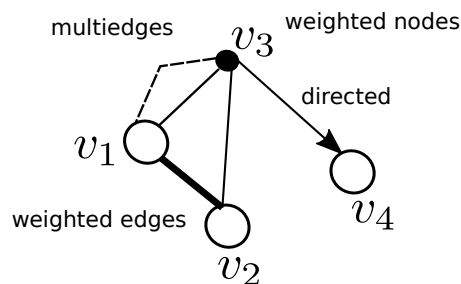


Figure 1.3: The complex networks may represent different system characteristics. The edges can be directed, weighted or multiply. Also, nodes can be assigned with different weights or any relevant feature.

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

Though the nodes in the same partition of a bipartite network are not directly connected, we can analyze their connections by projecting the bipartite network to one partition. The primary assumption is that two nodes in one partition could be connected if they point to the same node in the other partition. Figure 1.4 shows two projections of the bipartite networks. Consider the network of movies and actors. The one-mode projection of movies is an undirected network whose links indicate that two movies share the same actors. On the other hand, another projection is a network of actors. The links exist if two actors appear in the same movie [30, 53].

We should be aware that important information is lost when creating a one-mode projection. First, having weighted edges in the network of actors is necessary to know in how many movies two actors

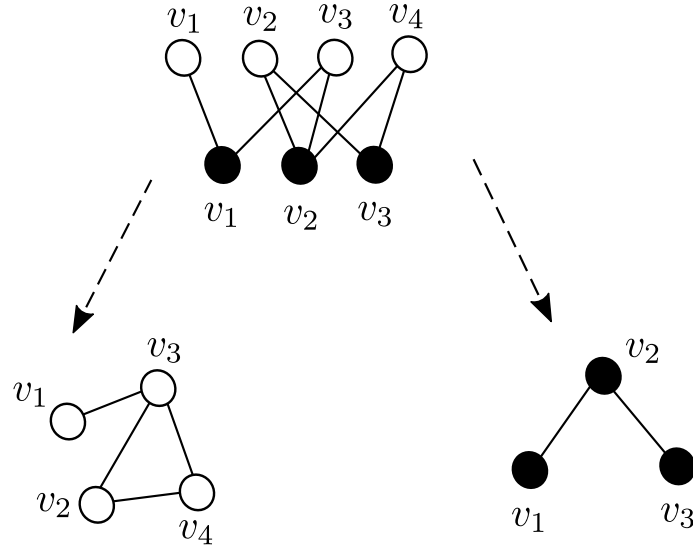


Figure 1.4: Bipartite network and two partition projections.

appear. From the one-mode projection, we can not reconstruct the original network. Moreover, two different bipartite networks may have the same projected networks. The important consequence of the network projection is the creation of cliques, i.e., subgraphs where all nodes are connected.

In general, it is possible to define the k -partite network. The same rules apply as before. There are k distinct node partitions, while the edges exist only between different types of nodes.

Temporal networks. Studying real systems as static networks can give us a lot of insight into the system's properties. Still, real systems are not static; they evolve not only in the number of elements but also in the number of interactions between them. Some interactions in the system may repeat in different intervals and could be described with complex activity patterns. Including time dimension in the network representation allows us to study the properties of the system closely. The temporal information may matter a lot [54]. For example, if the interaction between nodes (v_1, v_2) happened before in time than (v_2, v_3) , then nodes v_1, v_3 might not be connected, as is the case in the static network.

The temporal network is a collection of timestamped edges; as seen on Figure 1.5 - top panel. Each edge is defined as $(v_i, v_j, t, \Delta t)$, where v_i and v_j , are nodes t is time when interaction happen, and Δt is event duration [55]. The duration of the events may vary, as in the phone-call network. Also, for many systems, the time resolution of the event duration is too small. For example, this parameter may be neglected when people interact on social platforms or email each other because the event time is too short; it scales in seconds.

The temporal network can be represented as a sequence of static networks that evolve in time, $G = \{G(t_1), G(t_2), \dots, G(t_{max})\}$, as shown in Figure 1.5 - bottom panel. At each time step, we can create the network and analyze the macroscopic properties of the given network snapshot. With this, we can end up with graph snapshots with many disconnected components or empty graphs for some points [56]. Sometimes, a better approach is aggregating the links over time windows. Here, we need to specify the time window length w . Interactions in the time interval $0 \leq t < w$ enter the first snapshot. The following snapshot takes edges $w \leq t < 2w$, and so on. The time windows are not overlapping, but generally, it is possible to slide the time window for different periods $1 \leq \delta t < w$. The downside of this method is that we can not recover original data points. The larger the time window is, the more information is lost. If the time window is set to $w = t_{max}$, there is only one snapshot, and the temporal

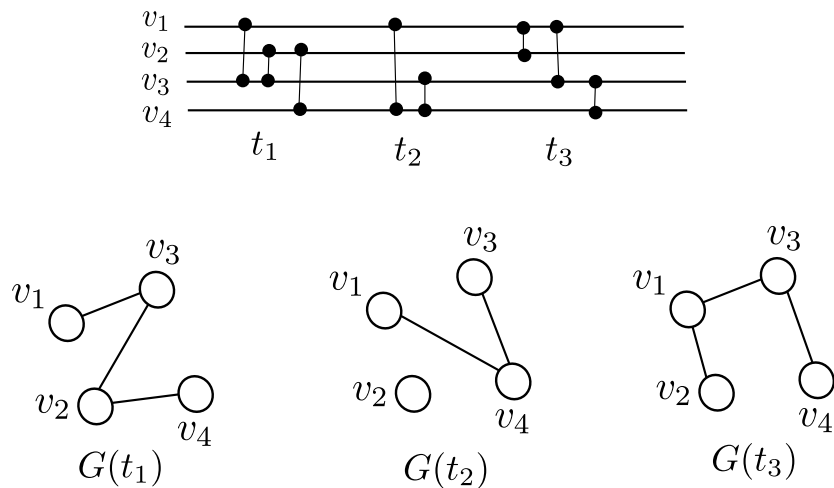


Figure 1.5: Top panel represents temporal network as collection of timestamped edges. Bottom panel represents sequence of static networks.

data are no more available [57, 58].

Multilayer networks were introduced for studying systems in which different types of interaction exist. This formalism allows one to investigate diverse network systems and combine different data types into one model [59]. In a multilayer or multiplex network, all nodes are present in each layer, but their interactions among layers differ. Two nodes may be connected in one layer but not in the other. Different online social systems may be an example of a multiplex network when users are connected on one platform but not on the other [60]. Another example is the airline transportation network, where each layer represents the flights of different airline companies [61].

1.2 Thesis outline

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

In chapter 2, we provide the methodology employed for this research. We describe the fundamental measures of complex networks and introduce basic complex network models. We review the most common probability distributions characterizing complex systems' properties and outline distribution fitting methods. Finally, we introduce the multifractality of the time series and dynamical reputation model.

Chapter 3 addresses the difference between network models where the growth in the number of nodes is constant and when it follows a non-trivial growth signal. This research aims to quantify how growth signals influence the structure of complex networks. Using the adapted aging model [62], we use computer simulations to generate different kinds of complex networks. For more realistic real-world network simulations, growing signals are time series of new users from online social platforms, MySpace, and Tech group from Meetup. They are described with trends, cycles, and long-range correlations. Often time series have multi-fractal properties. The results of this study are published in

[63], and they show the importance of growth signals in shaping the network structure because the scale-free networks, which represent real systems, are mainly altered.

As research on social groups mainly focuses on a single group, there are remaining questions about the characteristics of the entire system. For example, the Tech group is only one of the groups around which Meetup users organize; many other groups are created worldwide, so the system constantly grows. In chapter 4, we will examine how groups on online social platforms grow. The results are summarised in the paper [64]. This research is based on Reddit and Meetup data. From Meetup, we created two data sets, one with groups created in London and the other with groups created in New York, while for Reddit, we selected groups built before 2012. We are interested in explaining scaling behavior in group size and growth rate distributions and identifying the growth mechanisms present in the system. Using a bipartite complex network model, we can reproduce the universality found in the system.

Even though across complex systems, we find the emergence of universal behavior, for example, the scaling of the degree distribution of two groups is similar, different factors might influence its success. It is well known that many online groups may suddenly fall apart. These questions are the subject of the chapter 5, which main results are published in the paper [65]. Here, we study the question-answer platform Stack Exchange; it has more than 200 different topic-specific sites where people help each other answer questions. What is interesting about this system is that some sites were closed because they did not produce enough activity. For that reason, we selected the sites with the same topic that failed, but later, when someone proposed the site again, it stayed active. We analyze the evolution of user interaction networks; here, we use the temporal network approach and compare active and closed sites. We find that it is essential how the network users are distributed into a core-periphery structure [66]. The core must select firmly connected users, but their interaction with the periphery has to be high. In other words, a trustworthy core is needed to hold the community. Introducing the Dynamical Reputation Model (DIBRM) [67], based on user interaction sequences, we quantify how much users can be trusted and whether a community has a strong core. We briefly describe the Stack Exchange sites in the appendix A. In appendix B and C discuss how we choose parameters for the DIBRM model, while in appendix D we discuss the stability of inferred core-periphery structures.

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

Chapter 2

Methodology

2.1 The measures of complex network structure

The complex system can be represented by a complex network $G = (V, E)$, where the elements of a system (atoms, proteins, people) map to a set of N nodes $V = \{1, 2, \dots, N\}$. The interactions between elements map to L links between nodes, $E = \{e_1, e_2 \dots e_L\}$. There are a lot of measures to quantify the structure of the network. This section describes some of the important measures and their definitions on the undirected and unweighted networks, where the **adjacency matrix** $A = N \times N$ has value 1 if there is a connection between two nodes; otherwise, it is 0 [12]; as this network representation is mostly used through the thesis. We list degree distribution, correlations, and shortest path measures. We also discuss different structures found in the network, such as core-periphery or community structures.

2.1.1 Degree distribution

The simplest network measure is **node degree**, k . The degree of node i is the number of nodes adjacent to node i , $k_i = \sum_j A_{ij}$ [12, 30]. The network density is the average degree divided by $N - 1$, where N is the number of nodes [68].

In the case of regular networks, such as grids, each node has an equal degree, meaning that nodes in the network have similar roles. In the general case, the networks have a more complex structure. If the degree sequence is skewed, we can identify nodes with high-degree (hubs). Removing hubs may partition a connected network into several components [69].

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

Here we summarize the forms of degree distributions that are mostly found in the complex network theory:

- The Poisson distribution. The degree distribution in a random network, where all nodes have the same connecting probability, follows Poisson distribution $P(k) = \frac{(Np)^k e^{-Np}}{k!}$, where k is the mean degree distribution [53].

- Exponential distribution. $P(k) = e^{-k/k}$. It is the degree distribution of the growing random graph [53]. Even for infinite networks, all moments of distributions are finite and have a natural scale of the order of average degree.
- In many real networks, degree distribution follows a power law [53, 30]. $P(k) = k^{-\gamma}$, where γ is exponent of the distribution. No natural scale exists in this distribution, so they are called scale-free networks. In infinite networks, all higher moments diverge. If the average degree of scale-free networks is finite, then the γ exponent should be $\gamma > 2$. Therefore, real networks have a scale-free structure with the emergence of the hubs [30].

When plotting the degree distribution, it is common to use scaling of the axis. As many nodes have a low degree, like for power-law or exponential distribution, it is more useful to use a logarithmic scale [52]. Now it is easier to notice that data points follow a straight line, meaning that degree distribution is some exponential function.

2.1.2 Degree-degree correlations

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

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

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

Using the correlation coefficient definition, we can define correlations for vertex degrees [70]. For graph G which consists of n nodes and is characterized with with adjacency matrix \mathbf{A} and degree sequence $\mathbf{d} = [d_1, \dots, d_n]$, correlation of vertex degree has form:

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

An adjacency matrix allows us to calculate the correlations between neighboring nodes. If two nodes are not connected $A[i, j] = 0$, the degree of correlation between them does not contribute to the r .

The **degree-degree correlations** in the network are measured by **assortativity index**. If correlations are positive, networks are assortative; there is a tendency for connections to exist between similar degree nodes [53]. The negative correlations indicate that nodes with large degree are more likely to connect nodes with small degree, disassortative networks. The average first neighbor degree k_{nn} can be calculated as $k_{nn} = \sum_{k'} k' P(k'|k)$. The P is the conditional probability that an edge of degree k points to a node with degree k' . The norm is $\sum_{k'} P(k'|k) = 1$, and detailed balance conditions [12], $kP(k'|k)P(k) = k'P(k|k')P(k')$ [12]. If the node degrees are uncorrelated, k_{nn} does not depend on the degree; otherwise, increasing/decreasing function indicates positive/negative correlations in the network [71].

The Newman defined the assortativity [72] index r in slightly different way:

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

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

2.1.3 Clustering coefficient

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

$$c_i = 2e_i / (k_i(k_i - 1)), \quad (2.4)$$

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

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

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

2.1.4 Paths

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

The **path** between two nodes [70], v_i and v_j is a sequence of edges $\{(v_1, v_2), (v_2, v_3), \dots, (v_k, v_{k+1}), \dots, (v_{n-1}, v_n)\}$, where $v_1 = v_i$, $v_n = v_j$. In the path, the nodes are distinct. Otherwise, the sequence is called a **walk**, where each node can be visited many times. Also, it is possible to define a **cycle**, a path that starts and ends on the same node while other nodes in the cycle are distinct. The length of the path, walk or cycle is the number of links in the sequence. We can easily calculate the number of walks between two nodes using the adjacency matrix. The A^2 gives us walks of length 2, the A^3 , the number of walks of length 3, and so on.

The network is connected if it can define the path between every two nodes. When it is not the case, the network is disconnected into two or more connected components. Note that the component can be an isolated node. Also, in directed networks may happen that node v_i is reachable from node v_j , but if we start from v_j , we can not find the path to the v_i . Such a graph is connected but is called a weakly connected component [75].

We can find different paths between two nodes in the network, but the most important one is the **shortest path** [70, 75]. The distance between two nodes $d(v_i, v_j)$ is defined as the shortest path length between two nodes. In the case of weighted networks, it is the path with minimal weight, but its length is not necessary minimal. Distances on the network can give us insight into how similar networks are and indicate the node's relative importance in the network.

The **radius** is the minimum overall eccentricity value. In contrast, the **diameter** defines the largest distance between nodes in the network [70]. These definitions apply to directed and undirected graphs.

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

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

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

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

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

2.1.5 D-measure

For each node i , we can define the distribution of the shortest paths between node i and all other nodes in the network, $P_i = \{p_i(j)\}$, where $p_i(j)$ is the percent of nodes at a distance j from node i . The connectivity patterns can efficiently describe the difference between the two networks. To specify how much G and G' are similar we use D-measure [76]:

$$D(G, G') = \omega \left| \sqrt{\frac{J(P_1, \dots, P_N)}{\log(d)}} - \sqrt{\frac{J(P'_1, \dots, P'_N)}{\log(d')}} \right| + (1 - \omega) \sqrt{\frac{J(\mu_G, \mu_{G'})}{\log 2}}. \quad (2.7)$$

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

$$J(P_1, \dots, P_N) = \sum_{i,j} p_i(j) \log\left(\frac{p_i(j)}{\mu_j}\right), \quad (2.8)$$

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

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

2.2 Community structure

Nodes can be organized into groups called communities. In social networks, communities indicate that people share some common interests, or in biological networks, we can find that genes or neurons with similar functions are grouped. Identifying these hidden blocks can lead to interesting insights into the network. However, the community detection problem does not give a precise characterization of what a community is. A standard definition of a community is densely connected subgraph [77, 78], meaning that nodes in one community tend to associate, creating the assortative connectivity pattern. On the contrary, nodes could be organized in disassortative communities, where connections between groups are denser.

The network with k communities could be represented using $k \times k$ matrix p . The diagonal elements of p indicate the density inside communities, while off-diagonal elements show the density between groups. Figure 2.1 [79] shows the matrix and networks for two communities. In the first example, (2.1 a), the diagonal elements have a higher probability, as in the classic definition of assortative community structure. In disassortative structure (2.1 b), more connections exist between two partitions than inside them, i.e. off-diagonal elements have higher probabilities. Bipartite networks can be represented as a disassortative network with two groups. The links exist only between communities. Figure (2.1 c) shows the core-periphery network. This network structure is composed of a core where nodes are well connected with itself and with the periphery. The connectivity inside the periphery is sparse. Finally, if there is no difference between connectivity inside and between groups, the concept of communities is lost. We can treat the whole network as a single community, where each node has the same connectivity probability, i.e., as Erdos Renyi random graph.

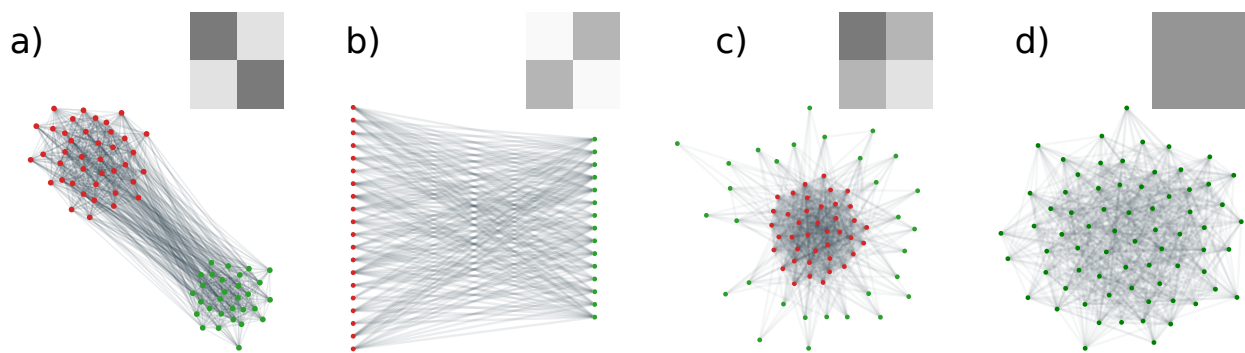


Figure 2.1: Different communities structures (a) assortative. (b) disassortative. (c) core-periphery. (d) Erdos Renyi random graph.

Different algorithms are used for detecting the community structure in the underlying network, optimizing different objective functions of the network partition. Still, if the ground-truth communities are unknown, there are no guarantees that we will infer the actual number of communities and entirely correct node assignments [80]. Even though community detection algorithms are widely used in complex network analysis as they can give us a better understanding of network structure [80, 81]. In this section are explained two community detection models, the first one based on optimizing the modularity function [77, 82], and the other based on the statistical inference of the Stochastic Block Model (SBM) where is optimized the likelihood function [77, 83, 84].

2.2.1 Community detection based on modularity function optimization

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

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

where the first part of equation measures number of links A_{ij} within community c , while second term is number of links within community if network is randomly connected $p_{ij} = \frac{k_i k_j}{2L}$. If the first term is larger than the second term, the modularity is positive and the partition is considered to be better than random, otherwise we can not consider that nodes in given group form community structure. The same idea can be generalized to the whole network: the modularity of the network partitioned into n_c communities is then defined as:

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

The higher modularity indicates that nodes are partitioned in better communities. When we put all nodes into only one community $M = 0$, otherwise, if each node is the community itself $L_c = 0$ and the sum is negative. The Newman showed that modularity function [88] applies for weighted networks.

Maximum network modularity indicates the best partions. As too many possible partitions exist, we need an algorithmic approach to identify the best separation. The first algorithm proposed for modularity optimization was **greedy algorithm**. First, it assigns each node to a community and starts with N communities. Then, we should merge each pair of communities and calculate the modularity difference ΔM . We can join those two communities by identifying the pair for which the difference is the largest. It is repeated until we get single community. The best partition is one with the largest M .

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

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

2.2.2 Stochastic block model

Another approach for studying the community structure of complex networks, the Stochastic Block Model (SBM), assumes that nodes are clustered in the groups, and the relations between nodes depend on the probabilities for group memberships [83]. In one group, nodes have similar connectivity patterns. To describe the network $G(N, L)$ with the SBM model, we need to define the following:

- k : number of groups.
- Group assignment vector, g : $g_i \in \{1, 2..k\}$, gives the group index of node i .
- SBM matrix, $p_{k \times k}$, whose elements p_{rs} are the probabilities that edges between groups r and s exist. Note that nodes within one group have the same connection probabilities.

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

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

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

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

The benefit of this model is that we can **generate** many networks with similar network structure [92]. When model parameters are initialized, the network can be easily generated. For each pair of nodes i and j in network G , we draw a link if random number $r_{ij} < p_{r,s}$.

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

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

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

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

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

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

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

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

Instead of generating networks, the opposite task is network **inference**. For a given network G , and specified the number of communities k , we can use the SBM model to infer the nodes' assignments into groups, so we need to choose vector g and SBM matrix p such that the likelihood for generating network G is maximized.

The formulation of the SBM model does not consider how to infer the optimal number of groups. Optimizing the likelihood function for different numbers of groups would increase likelihood while each node is not assigned to a different group. In practice, our found community structures for a fixed number of groups, and then the likelihood function could be penalized by the number of model parameters. One approach is calculating the **Minimum description length (MDL)** [84]. The variable which has probability $P(x)$, is described with amount of information $-\log_2 P(x)$. The numerator of posterior probability could be written as

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

where Σ is the data's description length (DL). The MDL consists of two terms: $\Sigma = -\log_2(p(G|p, g)) - \log_2 P(p, g)$. In the first part of the equation, the amount of information necessary to describe the model decreases with the number of groups [84]. The second contribution comes only from the model, and as the model becomes more complex, with a larger number of groups, this part increases [84]. The optimal solution represents the balance between these two terms in the MDL equation.

This SBM model has many variants motivated by specific properties of real data. For example, for degree heterogeneous networks, there is degree corrected SBM [93]. In some social networks, users can belong to more than one group, which can be modeled with mixed membership SBM. Other extensions include application to bipartite, weighted network, and hierarchical model [94]. Many community detection algorithms define the community as an assortative structure. With the SBM model, such limitations do not exist, and it is possible to directly use statistical inference for discovering core-periphery structures or even networks with bipartite structures.

2.2.3 Core-periphery structure

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

Another way to detect core-periphery structure is to use the inference method based on fits to a Stochastic Block Model (SBM) [98, 93]. In this method, we fit the observed network to a block model with two groups, such that edge probabilities have the form $p_{11} > p_{12} > p_{22}$. Vector $\theta_i = r$ indicates that node i is in block r , while SBM matrix $\{p\}_{2 \times 2}$, specify the probability p_{rs} that nodes from group r are connected to nodes in group s . The SBM model is looking for the most probable model that can reproduce a given network G [66]. Probability of having model parameters θ, p given network G is proportional to the likelihood of generating network G , prior of SBM matrix $P(p)$ and prior on block assignments $P(\theta)$: $P(\theta, p|G) = P(G|\theta, p)P(p)P(\theta)$, while the likelihood function takes following form: $P(G|\theta, p) = \prod_{i < j} p_{r_i s_j}^{A_{ij}} (1 - p_{r_i s_j})^{1 - A_{ij}}$, where A_{ij} is a number of edges between nodes i and j . The prior $P(p)$ is modified for core-periphery model such that $P(p) \sim I_{0 < p_{22} < p_{12} < p_{11} < 1}$, while prior $P(\theta)$ consists of three parts: probability of having 2 blocks; given the number of layers probability $P(n|2)$ of having groups of sizes n_1, n_2 and the probability $P(\theta|n)$ of having particular assignments of nodes to blocks.

2.3 The probability distributions

The shape of degree distribution is important for getting the first insight into the characteristics of the complex network. When nodes are generated randomly, and any two nodes are linked with the same probability p , we expect the binomial distribution. For larger networks it is Poisson distribution $P(k) = \frac{1}{k!} e^{-\langle k \rangle} \langle k \rangle^k$, where $\langle k \rangle = Np$. A different approach is to add one node and connect it randomly to the network at each time step. The obtained network then has the exponential degree distribution $P(k) = e^{-\lambda k}$. These are exponentially bounded distributions, meaning they decay exponentially or faster for the large values [53].

On the other hand, heavy-tailed distributions decay slower than exponential, and the events for large values are rare but still possible. For example, in the preferential attachment model, degree distribution emerges to the power law [53]. Also, many empirical data exhibit the heavy-tailed distribution. Even if they look like a power law, after statistical analysis, it may be concluded that the data deviate from the power law and could be equally good or even better fitted with some other distribution. Commonly used alternative distributions are lognormal distribution, stretched-exponential or power-law with an exponential cutoff.

This section gives an overview of relevant distributions and methods for fitting data and testing the quality of the performed fit. Figure 2.2 shows how different distributions look on linear (first column) and log-log scale (second column).

2.3.1 The properties of distributions

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

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

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

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

The power-law distribution is called scale-free distribution. If we scale the value k for the factor 2, the ratio of $p(x)/p(2x)$ is constant and does not depend on the k [52]. We'll find that these criteria are not satisfied by any other distribution

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

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

Lognormal distribution. The variable x has the lognormal distribution if the random variable $y = \ln(x)$ is distributed as normal distribution [101]

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

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

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

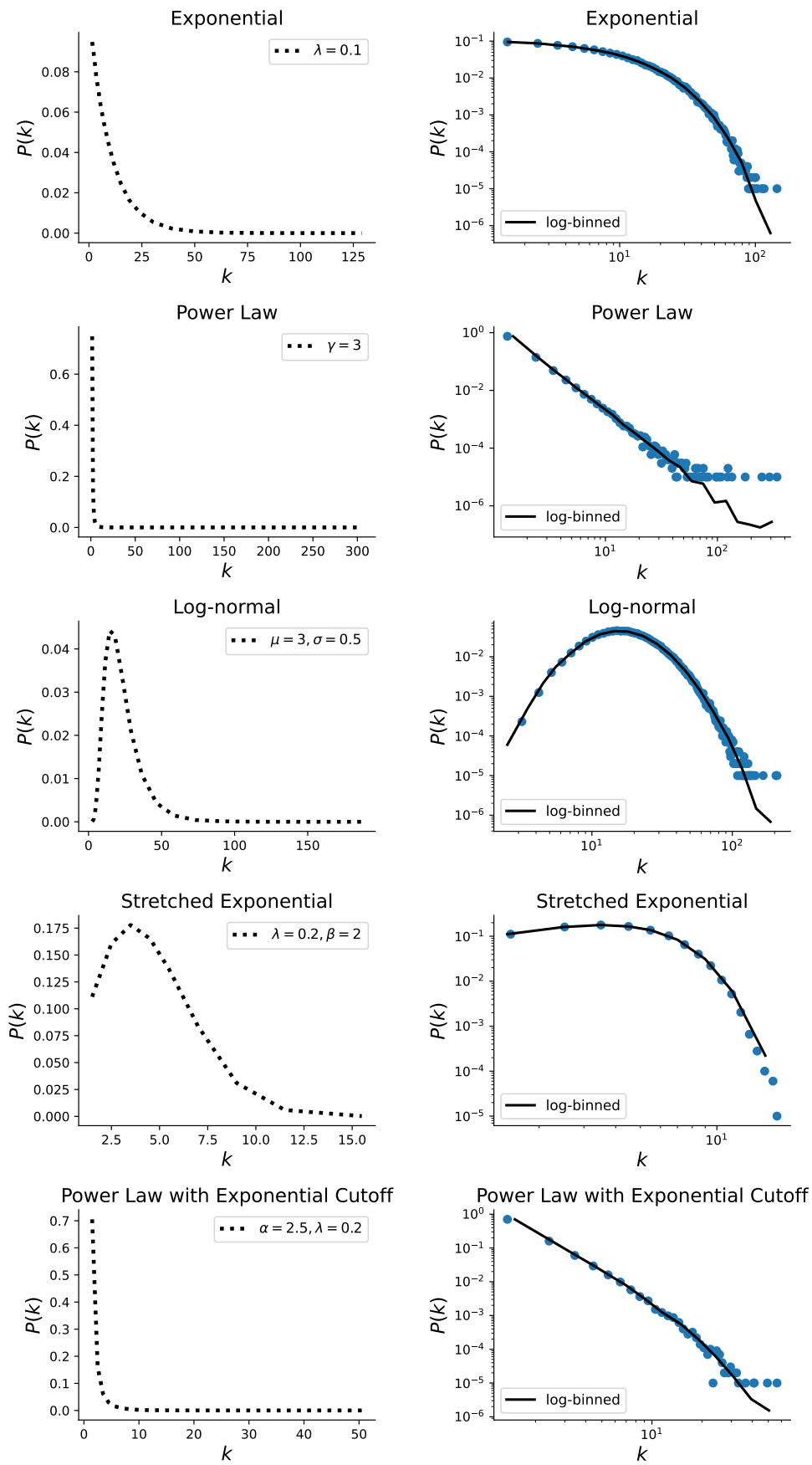


Figure 2.2: Probability distributions on a linear and double logarithmic scale.

The lognormal distribution has finite mean $e^{\mu+1/2\sigma^2}$, and the variance $e^{2\mu+\sigma^2}(e^{\sigma^2}-1)$. [99]. Despite the finite moments, the lognormal distribution can be similar to the power-law distribution. If the variance is large, then the probability function on the log-log plot appears linear for a large range of values.

Using the **multiplicative processes**, we can generate the lognormal distribution [52, 99]. The lognormal distribution is generated by processes that economist Gibrat called the law of proportionate effect. If we start from the organism of size S_0 , at each time step, the organism may grow or shrink according to the random variable ϵ [99]

$$S_t = \epsilon_t S_{t-1}. \quad (2.22)$$

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

$$S_t = \epsilon_t S_{t-1} = \epsilon_t \epsilon_{t-1} \dots \epsilon_2 \epsilon_1 S_0. \quad (2.23)$$

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

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

The sum of the logarithms of the ϵ_t , according to the Central Limit Theorem (CLT), follows the normal distribution. The CLT states that the sum of identically distributed random variables with finite variance converges to the normal distribution. If $\ln(S_t)$ is normally distributed, then S_t follows the lognormal distribution [99].

The multiplicative processes generate the lognormal distribution. Introducing a threshold in the multiplicative process leads to the power law. For example, in the Champernowne model [52], individuals are divided into classes according to their income. The minimum income is m . People between incomes m and γm are in the first class, and the second class is people with incomes between γm and $\gamma^2 m$. The individuals can change their class, so it is described as a multiplicative process, but with a threshold, as income can not be lower than m . If we fix $\gamma = 2$, and consider that with probability $p_{i,i-1} = 2/3$, the change is from higher to lower class. In contrast, with probability, $p_{i,i+1} = 1/3$ individual goes to a higher class. In this process, the distribution of incomes emerges as the power-law distribution.

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

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

where $k > 0$ and $\gamma > 0$. This function combines the power-law and exponential terms responsible for an exponentially bounded tail [53]. Taking the logarithm $\ln(p(k)) = \ln C - \gamma \ln k - \lambda k$, when $k \ll 1/\lambda$ the second term dominates, so distribution follows the power-law, with exponent γ . Otherwise, the λx term dominates, resulting in an exponential cutoff for high values.

Stretched exponential The stretched exponential distribution is defined as:

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

the parameter β is stretching exponent determining the properties of the function $p(k)$ [53]. For $\beta = 1$, the function is exponential. For $\beta < 1$, it is hard to distinguish the distribution from the power law. We have a compressed exponential function for $\beta > 1$, so k varies in the narrow range.

2.3.2 Estimating the distribution parameters

The maximum likelihood estimation (MLE) is a method where we consider that data comes from a particular distribution, so we want to maximize the likelihood of the data to find the distribution parameters. For a given set of i.i.d. observations x_1, x_2, \dots, x_n , sampled from the distribution $p(x)$, we can define the likelihood function [102]. The likelihood function tells us how likely it is to have the given data if the distribution parameters are θ

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

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

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

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

First, we fit empirical data to get model parameters and calculate the KS statistics of this fit [103]. Then, many synthetic data sets are generated with model-optimized model parameters. Then each synthetic data set is fitted, and KS statistics are obtained relative to its model. From there, we can calculate **p-value**, the fraction of times that KS-statistics in synthetic distributions is larger than in empirical data. If $p - \text{value} < 0.1$, we reject the hypothesis that this distribution describes the empirical data. Otherwise, the model can not be rejected. Failing to reject the hypothesis does not mean the model is a correct distribution for the data. Other distributions might fit the data equally good or even better. To have an accurate p-value, we need a large sample. For a small number of synthetic distributions, it is possible to have a high p-value, even if the distribution is the wrong model for the data. Finally, we need to be confident in obtained results. The same procedure can be repeated for different distributions. If the p-value for the power law is high, while for alternative distribution, it is low, we can conclude that the power law is a more probable fit.

Another method, the **likelihood ratio test**, allows us to compare two distributions directly [103]. The distribution with a higher likelihood under empirical data is a better fit. We can calculate the likelihood ratio, or it is easier to obtain the likelihood ratio's logarithm because its sign determines which distribution is a better fit. For given two distributions $p_1(x)$ and $p_2(x)$.

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

$$\mathcal{R} = \sum_{i=1}^n [\log p_1(x_i) - \log p_2(x_i)]. \quad (2.28)$$

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

$$\sigma^2 = \frac{1}{n} \sum_1^n [(l_i - l_i) - (\langle l \rangle^{(1)} - \langle l \rangle^{(2)})].$$

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

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

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

Here we use the standard two-tail hypothesis test [103], assuming that the null hypothesis is $R = 0$. If the p-value is larger than a threshold, the R sign is unreliable, and the test does not favor any distribution. If p is small, $p < 0.1$, then it is unlikely that the observed sign is obtained by chance, so we reject the null hypothesis that $R = 0$.

2.4 Network models

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

2.4.1 Random network model

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

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

- We start with N isolated nodes.

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

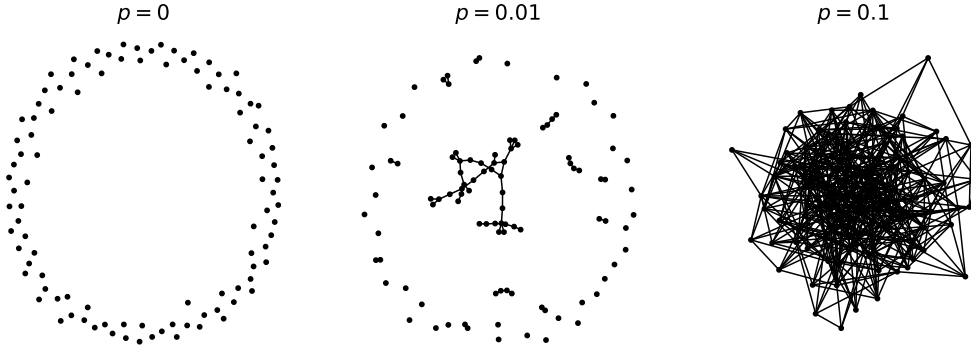


Figure 2.3: Erdős-Rényi graph with $N = 100$ nodes and different linking probabilities p .

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

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

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

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

$$\langle k \rangle = (N - 1)p. \quad (2.31)$$

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

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

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

The binomial distribution describes very well small networks, see Figure 2.4. For larger networks, we find that they are sparse and that the average degree is much smaller than a number of nodes $\langle k \rangle \ll N$. In this limit, binomial distribution becomes the Poisson, as could be shown in Figure 2.4, which now depends only on one parameter $\langle k \rangle$

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

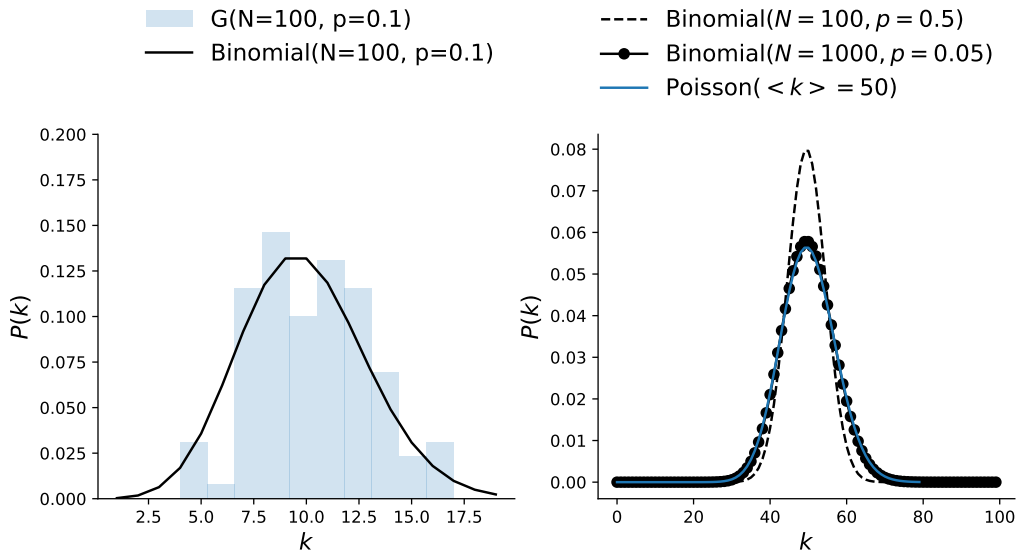


Figure 2.4: Degree distribution of ER graph. The degree distribution of small networks follows binomial. Larger networks are better approximated with Poisson distribution, and degree distribution for fixed average degree $\langle k \rangle$ becomes independent of the network size.

The random graph has a very small **average path length**, it is given as $\langle l \rangle = \frac{\ln N}{\ln(pN)}$ that is characteristic of many large networks [104]. The clustering coefficient is proportional to linking probability, $\langle C \rangle = p$, so we find a small clustering coefficient in large random networks, contrary to real-world networks.

Figure 2.3 shows how the network becomes more connected by increasing the linking probability p . When $p = 0$, all nodes are disconnected. In the other limit, $p = 1$, the network is fully connected. Between those two probabilities exists critical probability, where the giant component appears. The giant component is a sub-graph whose size is proportional to the network size. In other words, the network does not have disconnected components. Such change in the network is a phase transition in network connectivity and is related to percolation theory.

The phase transition occurs when the average degree is $\langle k \rangle = 1$, which gives us: $p_c = \frac{1}{N-1}$, meaning that all nodes have degree larger than one [53]. When the $\langle k \rangle < 1$, the network is in the 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.

2.4.2 Small-world networks

Inspired by the idea that real-world networks are highly clustered and the average distance is small, Watts and Strogatz [31] proposed the "small-world" model. The model starts from the regular lattice, and with rewiring links, the network starts to resemble small-world property. The procedure is the following:

- At the beginning, nodes are placed on the ring lattice, see Figure 2.5, and each node is connected to $k/2$ first neighbors on the left and the right side. Initially, the clustering coefficient is high, $c = 3/4$.
- For each link in the network, with probability p , we choose a random node to rewire the link. This connects long-distance nodes, decreasing the network's average path length, Figure 2.5.

The model interpolates between the regular graph when the probability is $p = 0$ and the random graph with $p = 1$ when all links are randomly rewired. Short distances and high clustering are present in the network for the relatively small probabilities ranging from $p \approx 0.01 - 0.1$ [31].

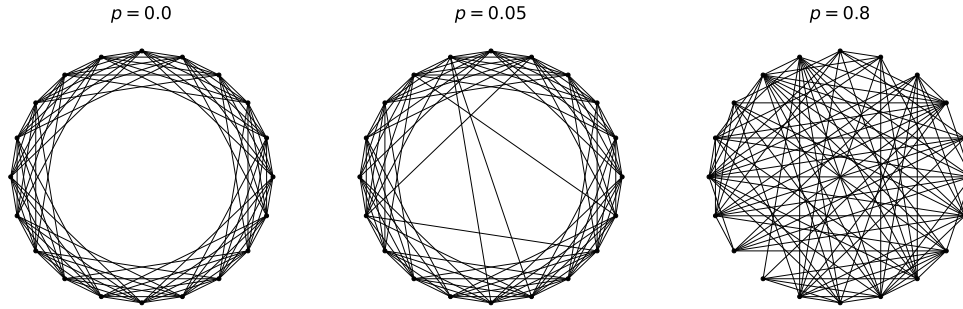


Figure 2.5: Watts and Strogatz graph model creation, for different rewiring probabilities.

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

2.4.3 Barabási-Albert model

The ER random graph model and WS small-world model are static models where the number of nodes is fixed. It is one of the reasons why they can not fully explain the properties of real systems. The size of real systems does not remain constant; real networks grow. Growth means that at each time step, new nodes are added to the network. The simplest model that produces scale-free networks is the Barabasi-Albert model [32].

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

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

The simplest formulation of the preferential attachment model is that new nodes tend to connect with high-degree nodes. The linking probability Π is then proportional to node degree k [105]

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

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

First, we can calculate the evolution of network degrees in time [105].

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

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

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

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

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

Degree distribution follows power-law, and for large k is approximated with $P(k) = k^{-\gamma}$, where $\gamma = 3$. As the network grows, nodes with larger degrees become bigger, and we end up with few nodes with many links, called hubs. Figure 2.6 - left pane shows generated BA network, consisting of $N = 100$ nodes, where even on this scale, we can notice the emergence of hubs. The right pane of Figure 2.6 shows obtained degree distribution of a larger network with $N = 10^4$ nodes. The degree distribution is also independent of the time and size of the system, meaning the emergence of a stationary scale-free state. If we vary m , the slope of distributions is the same, but they are parallel. After rescaling $p(k)/m^2$, they fall on the same line [53].

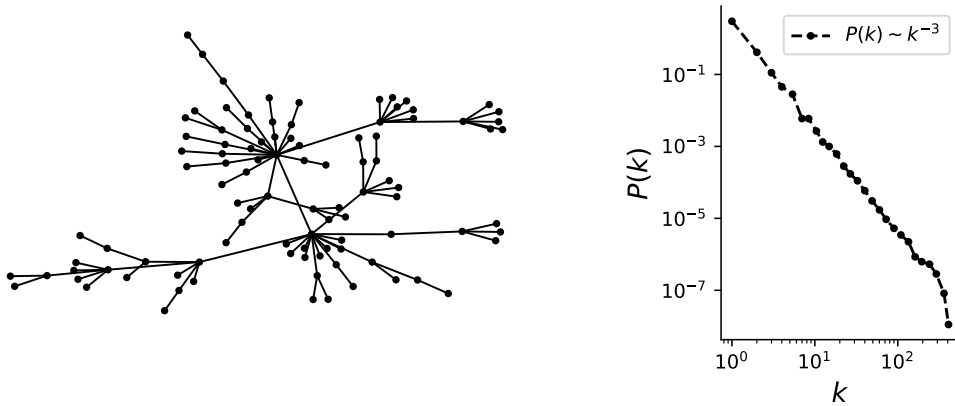


Figure 2.6: Barabasi-Alber model. The left panel shows the BA network, with 100 nodes. The right panel shows the degree distribution for BA network of 10^4 nodes that follow the power-law.

The **network diameter**, represents the maximum distance in network, $d \sim \frac{\ln N}{\ln \ln N}$ [104]. The diameter grows slower than $\ln N$, making the distances in the BA model smaller than in the random graph. The difference is found for large N . It is known that the BA network has hubs that shorten the path between less connected nodes. Also, if hubs are removed from the network, the network easily partitions into several components, losing its properties. The **clustering coefficient** of the BA model follows $C \sim \frac{\ln N^2}{N}$ [104]. It differs from clustering found in random networks, and BA networks are generally more clustered.

The combination of the growth and preferential attachment linking is crucial for getting scale-free networks [32]. For example, eliminating the preferential attachment; in a growing network with random linking, degree distribution is stationary but follows exponential. In contrast, the absence of growth leads to the non-stationary degree distribution. When a number of nodes is fixed, the network

grows only in the number of links, such that randomly chosen node i connects to node j according to probability Π . In the beginning, the degree distribution follows the power law, the same as in the BA model. As more links are added to the network, the distribution changes its shape; first, the peak appears, while at the end network becomes a complete graph, where all nodes have the same degree.

2.4.4 Nonlinear preferential attachment model

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

$$\Pi(k_i) = k_i^\beta. \quad (2.38)$$

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

For $\beta > 1$, preferential attachment effects are increased, leading to super hubs' emergence. The hub-and-spoke network appears in this regime, where almost all nodes are connected to a few high-degree nodes [106].

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

2.4.5 Aging model

To understand how aging can impact the network structure, we look into probability dependent on two parameters, nodes degree k and age of node i at the time point t $\tau_i = (t - t_i)$, where t_i is the time when node i is added to the network [33]

$$\Pi_i(t) \sim k_i \tau_i^\alpha. \quad (2.39)$$

The parameter α controls the linking probability dependence on the nodes' age, as could be seen on Figure 2.7. If $\alpha = 0$, the aging of nodes is disregarded. If $\alpha > 0$ is positive, the older nodes are more likely to create connections. In this regime, the preferential attachment stays present, and the high-degree and older nodes are preferred. For very high α , each node is connected to the oldest node in the network. The scale-free properties are present; the power-law exponent γ deviates from $\gamma = 3$. It is found that γ ranges between 2 and 3. When α is negative, aging overcomes the role of preferential attachment, and scale-free properties are lost. For significant negative α network becomes a chain; the youngest nodes are those who get connected.

In the general aging model, the non-linearity on the node degree is introduced, so this model has two tunable parameters α and β . The probability that a link is created between the new node and the existing node is defined as [62]

$$\Pi_i(t) \sim k_i(t)^\beta \tau_i^\alpha. \quad (2.40)$$

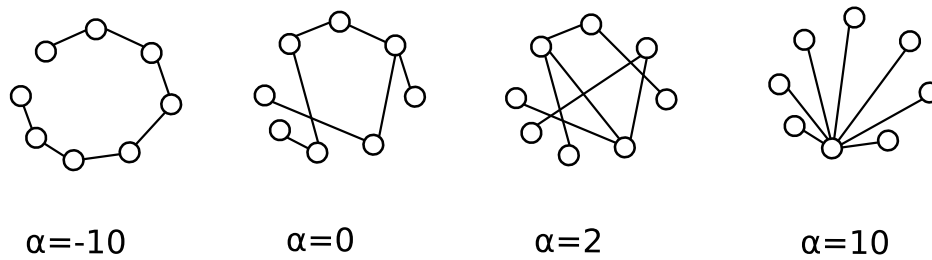


Figure 2.7: Dependence of parameter α and network structure. Network topology vary from chain network to the case where each node is connected to youngest node.

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

- For example if we fix $\beta = 1$ and $\alpha = 0$ generated networks are scale-free; degree distribution is $P(k) \sim k^{-\gamma}$ with $\gamma = 3$.
- In the case of nonlinear preferential attachment $\beta \neq 1$ and $\alpha = 0$ scale-free properties disappear.
- Scale-free property can be produced along the critical line $\beta(\alpha^*)$ in the $\alpha - \beta$ phase diagram, see Figure 2.8.
- For $\alpha > \alpha^*$ networks have **gel-like small world** behavior.
- For $\alpha < \alpha^*$ and near critical line $\beta(\alpha^*)$ degree distribution has **stretched exponential** shape.

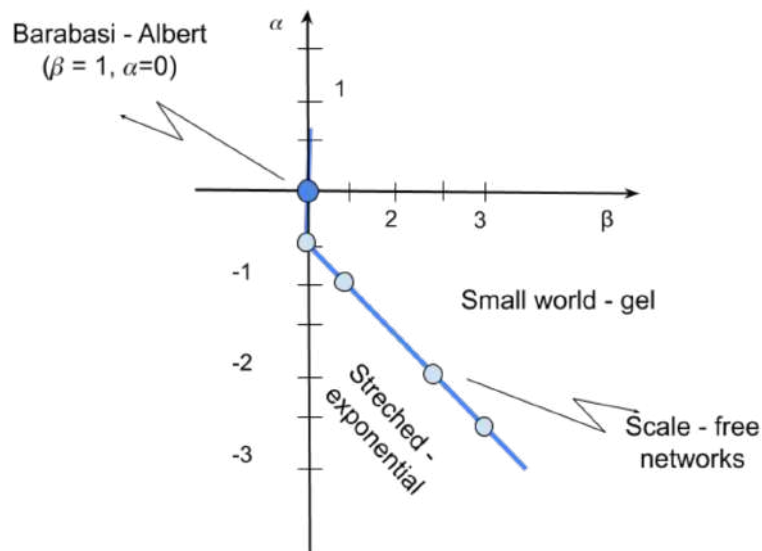


Figure 2.8: Phase diagram of aging network model.

2.5 Fractal analysis

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

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

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

2.5.1 Long and short-term correlations

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

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

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

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

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

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

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

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

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

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

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

2.5.2 Rescaled range analysis

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

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

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

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

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

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

2.5.3 Fluctuation analysis

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

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

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

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

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

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

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

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

2. Methodology

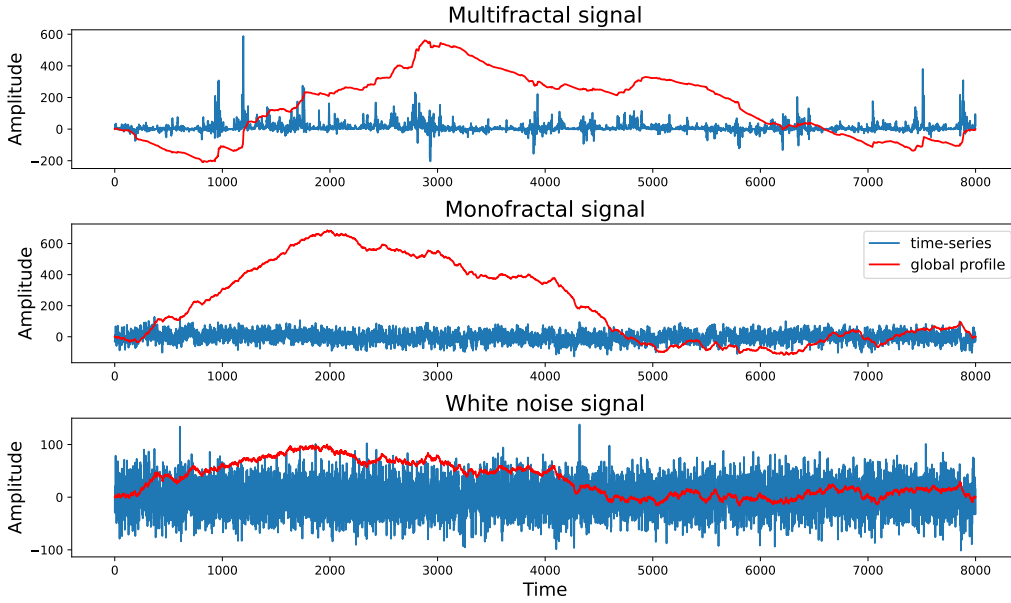


Figure 2.9: Multifractal, monofractal and white noise signals.

The trends present in the time series do not have to be linear [111]. The middle and bottom panel in Figure 2.10 show that the segments of the signal could be very well approximated with some higher order functions: quadratic or cubic. In general, using the detrended fluctuation analysis (DFA) we could remove the polynomial trend of the order m [112]. From each segment ν , local trend $p_{\nu,s}^m$ - polynomial of order m - should be eliminated, and the variance $F^2(\nu, s)$ of a detrended signal is calculated as in equation:

$$F^2(\nu, s) = \frac{1}{s} \sum_{j=1}^s [Y(j) - p_{\nu,s}^m(j)]^2. \quad (2.44)$$

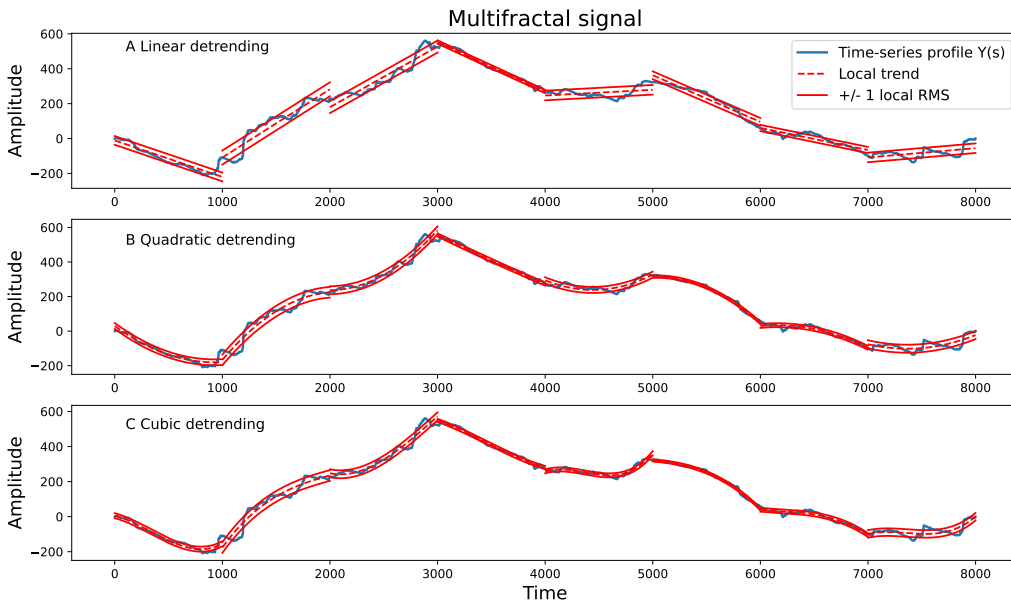


Figure 2.10: Detrending of multifractal signal for the segments of length $s = 1000$. Panel A- linear detrending, panel B- quadratic detrending, panel C- cubic detrending.

2.5.4 Multifractality of the signals

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

Multifractal detrended fluctuation analysis (MFDFA) is used [113, 114] to estimate multifractal Hurst exponent $H(q)$

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

The MFDFA for $q = 2$ is equivalent to the DFA method. The value of $H(0)$, which corresponds to the limit $F(q), q \rightarrow 0$, cannot be calculated directly because the exponent diverges. Instead, the logarithmic averaging procedure has to be considered.

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

The fluctuating function scales as power-law $F_q(s) \sim s^{H(q)}$ and the analysis of log-log plots $F_q(s)$ gives us an estimate of multifractal Hurst exponent $H(q)$, see Figure 2.11.

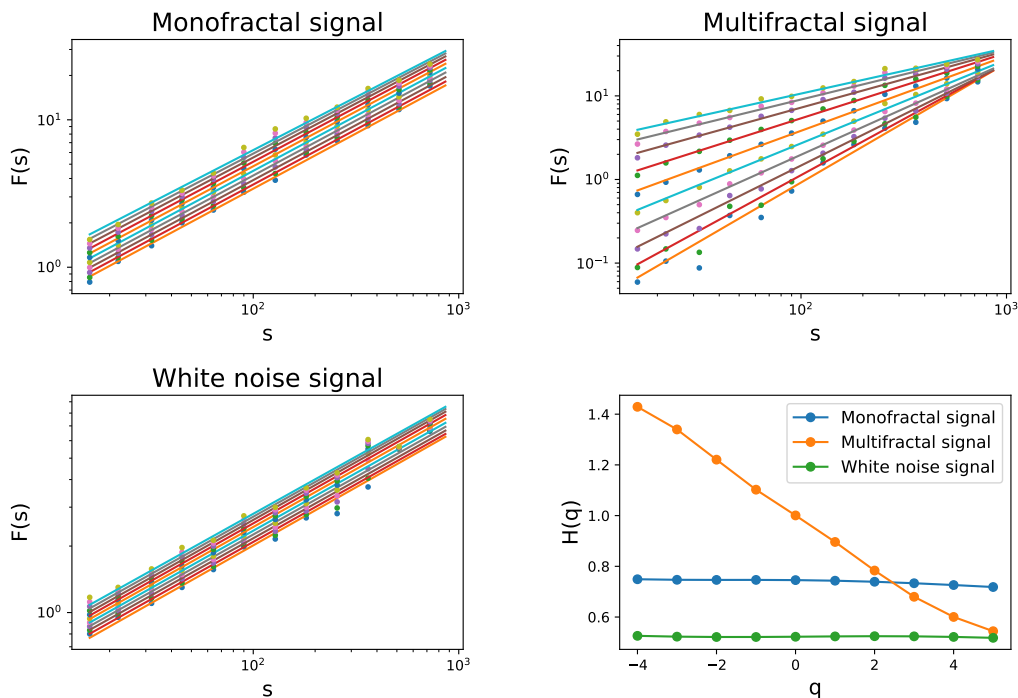


Figure 2.11: Dependence of the fluctuating functions on the scale for monofractal, multifractal and white noise signals, and the Dependence of the Hurst exponent H on the scale $1/q$ for different types of signal (bottom right).

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

2.6 Dynamical reputation model

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

With Dynamic Interaction Based Reputation Model (DIBRM) [67], we can quantify the user reputation R_n after each interaction using equation 2.46, where n is the number of interaction $n \in 1, N$

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

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

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a}, \quad (2.47)$$

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

$$I_n = I_{bn} \left(1 + \alpha \left(1 - \frac{1}{A_n + 1}\right)\right). \quad (2.48)$$

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

For example, if we set the characteristic window size and basic value of interaction to $t_a = 1day$, $I_{bn} = 1$, we can analyze the influence of the parameters α and β on the user reputation. Lower α and β values lead to faster reputation decline, as shown in Figure 2.12 - left panel. With lower β , the reputation may quickly drop close to the reputation threshold, under which we don't consider the user as active. In contrast, with larger values of β , reputation stays high even if a user is inactive for a larger period. The parameter α is the most important influence on burst behavior, where larger α leads to higher reputation values.

If a user is frequently active, we can record the reputation after each day. On the other hand, if $t_n - t_{n-1} > 1day$ we need to interpolate the reputation values for each day between two interactions,

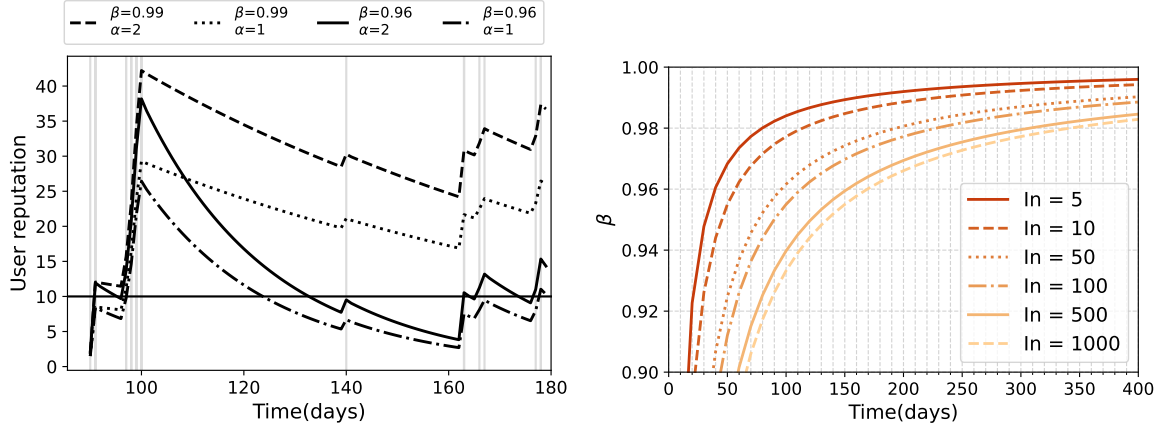


Figure 2.12: Left panel shows the dynamics of user reputation obtained in DIBRM model for different model parameters α and β . Right panel shows the dependence of parameter β and number of days to reputation from starting value I_n drops below threshold $I_n = 1$.

$t_{n-1} < t_d < t_n$. To do that, we consider that due to inactivity, reputation will only decay, so it could be calculated as $R_d = R_{n-1}\beta^{\Delta_d}$, where $\Delta_d = (t_d - t_{n-1})/t_a$.

When a user becomes inactive, its reputation starts to decline, and when it drops below the reputation threshold user does not have any influence on the community. We can approximate the dependence of parameter β and time δt needed for reputation to reach this level as $\beta = \left(\frac{R_0}{R_t}\right)^{\frac{t_a}{\delta t}}$. In the examples in Figure 2.12, - right panel, the parameter $t_a = 1$ day, while we vary different starting reputation levels I_n . For β values below 0.96, the decay is fast, and within two to four months of inactivity, even high reputation values are reduced below the threshold. On the other hand, with values of β , the decay process is more differentiated, and the high reputation becomes harder to lose, surviving up to a year of inactivity. For β equal to 0.96, reputation with starting value 5 needs around one month to decay below the threshold. For higher reputations, 500 or 1000, the decay period is around 5 months.

In this model, the user's reputation changes continuously through time, decreases when the user is inactive, and grows with frequent and constant user contribution. The reputation has highest growth when user shows burst in activity. With model parameters, $I_{bn}, t_a, \alpha, \beta$, the dynamic of user reputation may be controlled and adapted to different communities. If the community has its reputation system, we can also fit the model parameters to mimic the actual reputation dynamic. In this thesis DIBRM model is used to analyze Stack Exchange communities, Chapter 5, while in Appendix B, we suggest the procedure to estimate the model parameters for this specific system.

Chapter 3

Evolving complex network structure dependence on the properties of growth signals

Complex networks grow by adding new nodes, and growing network models consider growth constant over time. This approximation is sufficient for explaining how properties of complex networks can emerge; for example, we find power-law degree distribution in the Barabasi-Albert model [32]. Models mainly focus on linking rules and their influence on the topology of complex networks.

Still, the growth of real systems changes over time. In online social networks, new users join daily, and the users' activity might have bursty nature. We can consider a co-authorship network, where links are created between scientists when they publish a paper [115, 116]. The dynamics of real networks can be complex and highly influenced by nonlinear signals. The growth signal, the number of new nodes in each time step, has cycles and trends. Circadian cycles are directly reflected in growth signals, and we also find long-range correlations and multifractal properties [108].

In this chapter, we study how growth signals influence the network structure. We explain the properties of growth signals, both real and computer-generated, and analyze networks created with a growing network model where the interplay between aging and preferential attachment shapes their structure. We are interested in incorporating non-constant growth signals into the model and measuring their impact on complex networks. Differences between networks with the same number of nodes and links can be observed by analyzing connectivity patterns. Figure 3.1 summarizes our goals.

3.1 Aging network model with growth signal

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

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

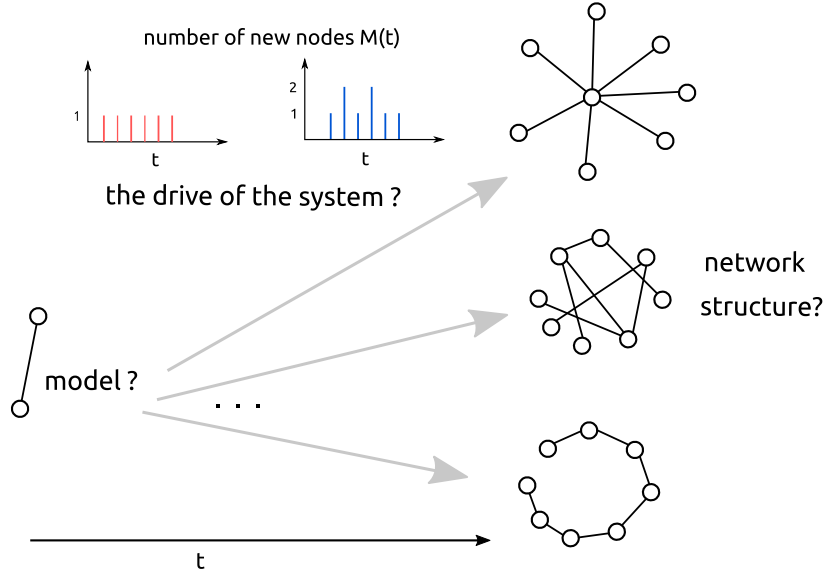


Figure 3.1: The open question is how nonlinear signals, in combination with the network model, influence the network’s structure. Under what circumstances do networks have the scale-free, hub-spoke, or chain structure?

We add $M(t)$ nodes with L links at each time step. As multiple links between two nodes are not allowed, we’ll get $M(t)$ new nodes with degree L , which describes the third term in the equation. Old nodes can increase their degree from 1 to $M(t)$, as different new nodes can choose the same node. The first term in the equation describes nodes with degree $k \in \{k - M(t), \dots, k - 1\}$ that getting degree k , while in second term nodes with degree k entering degree $k \in \{k + 1, \dots, k + M(t)\}$. The quantities $r_{k-j \rightarrow k}$ and $r_{k \rightarrow k+j}$ are the rates that express the transitions of a node from class with degree $k - j$ to one with degree k and from class with degree k to class with degree $k + j$ respectively.

For the model, we choose the aging model where linking probability depends on network degree k and its age τ , $\Pi_i(t) \sim k_i(t)^\beta \tau_i^\alpha$. With this linking probability, the master equation was solved for $M(t) = \text{const.} = 1$, using approach [34]. When $M(t)$ is the correlated function, the equation is not solvable analytically. Instead, we use numerical simulations to study the influence of the signal $M(t)$ on the network structure. When we add only one link per node $L = 1$, networks are uncorrelated trees. To obtain the clustered structures, we need to use $L > 1$; each new node can create more than one link. Finally, we focus on the aging model parameters $-\infty < \alpha \leq -1$ and $\beta \geq 1$. We expect a critical line $\beta(\alpha^*)$ where scale-free networks can be found. Under critical line, networks have stretched exponential degree distribution, and for large β small-world networks are present.

Finally, we need to define the new nodes’ time series. We focus on the growth of two real systems, the **TECH** [117] community in the Meetup website and on two months of **MySpace** [118] social network. Besides these signals, we use randomized MySpace and TECH signals and uncorrelated Poissonian signals.

3.1.1 Characteristics of growth signals

MySpace signal is the number of new members who appear for the first time in the data. Here, the time step is one minute. The MySpace signal has $T = 3162$ steps, with $N = 10000$ members. To describe the properties of the signal, we use Multifractal detrended analysis and calculate the Hurst exponent on different scales, showing the right pane of the Figure, 3.2. It is multifractal $q < 0$ and

becomes constant for $q > 0$; it has long-range correlations as $H(q = 2) = 0.6$. My Space signal has cycles characteristic of the human circadian rhythm, Figure 3.2. We can easily destroy trends and cycles if we randomize the MySpace signal. The randomization is done with the reshuffling procedure, where we keep the number of nodes, length, and the mean value of the signal. The inset of the original and randomized signals show the time series' global profile; we find that trends are destroyed. Also, the randomized MySpace signal no longer has long-range correlations; the Hurst exponent indicates short-range correlations $H = 0.5$, and the signal becomes monofractal.

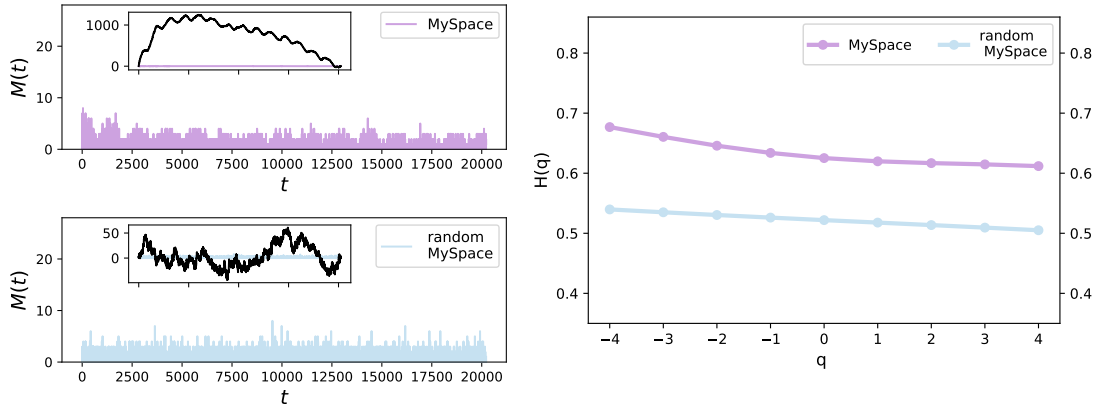


Figure 3.2: MySpace signal, the random MySpace signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q (right pane).

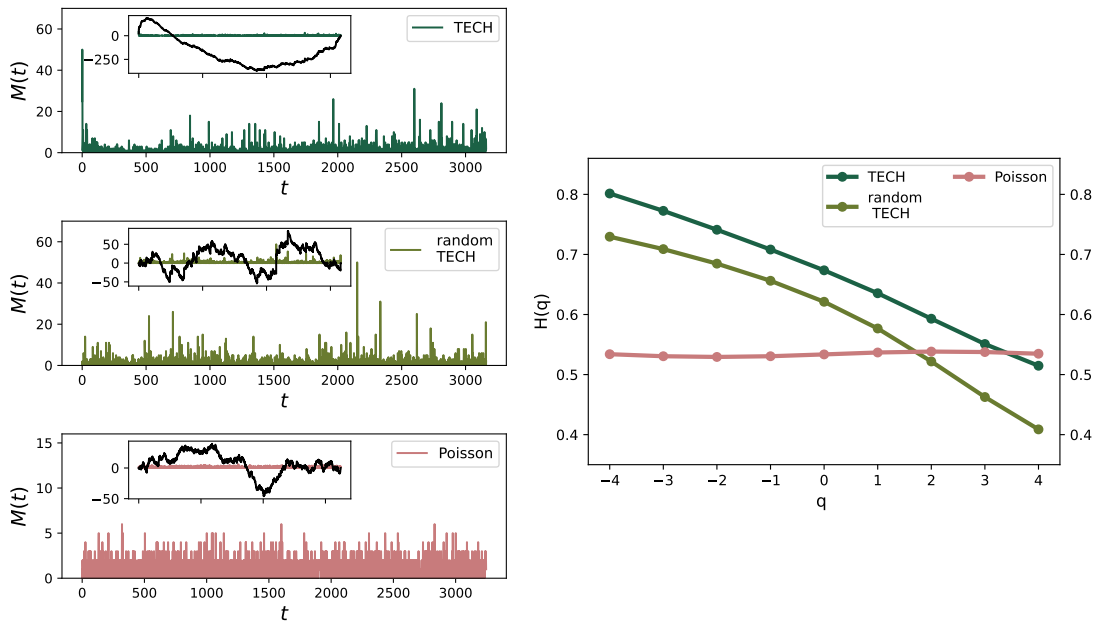


Figure 3.3: TECH signal, the random TECH signal (left pane) and the dependence of multifractal Hurst exponent $H(q)$ of the scale q (right pane).

The TECH is a group from the Meetup website that gathers users interested in technology. Using the Meetup website, they organize offline events. The time unit in this time series is an event since then are created links between events. The TECH time series $M(t)$ represents the number of users who joined the TECH community and visited the event for the first time. The time series length is $T = 3162$ steps, and we count $N = 3217$ members in the TECH community for a given period, Figure 3.3. TECH signal has long-range correlations with Hurst exponent $H(q = 2) = 0.6$. Also, we find that TECH

is multifractal, as the Hurst exponent is not constant across the scales. The multifractality originates not only from signal trends but also from the broad probability distribution of time series. If we randomize the TECH signal, we can easily destroy trends and cycles, but the signal keeps multifractal properties, meaning that broad probability distribution can not be eliminated. Therefore, we generate the uncorrelated signal from the Poissonian probability distribution. The length of this signal is $T = 3246$, while we keep the number of nodes N the same as in the TECH signal.

3.1.2 Structural differences between evolving complex networks

We can compare the networks with the same number of nodes and links generated with growth signals with different properties. We use a growing network model where we vary parameters $-3 < \alpha \leq -1$ and $-3 \leq \beta \leq 1$. We also vary the network density, $L \in \{1, 2, 3\}$. For each set of model parameters α, β, L and each signal $M(t)$, we create the sample of 100 networks. Besides this, for the same set of parameters, we generate the sample of networks with $N = 10000$ and $N = 3217$ nodes grown with constant signal $M(t) = 1$; one node is added to the network at each time step. To examine how different growing signals influence the structure of networks, we use D-measure [76], defined methodology chapter. We equally consider the global and local properties, setting parameter $w = 0.5$. We compare the networks grown with the constant and fluctuating signal with D-measure for all network pairs between two samples and average the result. The advantage this measure has is that it can measure the distance between two network structures, even if they are generated with the same model; that was not the case with Hamming distance or graph editing distance [76].

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

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

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

3.1.3 The structure of networks

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

Figure 3.5 shows properties of networks generated with model parameters $L = 2, \alpha = -1.0, \beta =$

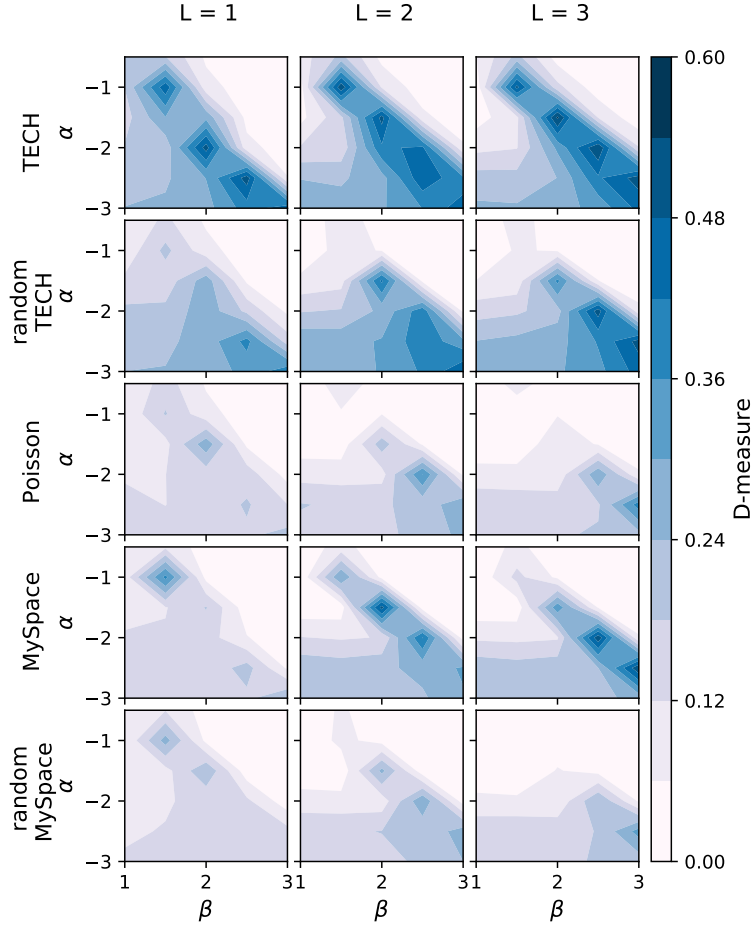


Figure 3.4: The comparison of networks grown with growth signals shown in figures 3.3 and 3.2 versus ones grown with constant signal $M = 1$, for the value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in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

1.5, that lie on the critical line. The degree distributions $P(k)$ of networks generated with real signals TECH and MySpace have super-hubs emerged. Degree distributions generated with randomized and white noise signals do not differ from the degree distribution of networks generated with the constant signal. Networks generated with real signals average neighboring degree $\langle k \rangle_{nn}(k)$ and clustering coefficient $c(k)$ depend on node degree. In contrast, networks generated with constant and randomized signals weakly depend on the degree k .

We also find structural differences between networks, obtained with model parameters under the critical line $\alpha < \alpha^*$, see Figure 3.5. The difference is mainly found in the TECH signal. Degree distribution $P(k)$ shows the emergence of hubs in networks grown with TECH signal, while the randomized and Poisson signals are more similar to networks grown with the constant signal. MySpace signal, whose generalized Hurst exponent $H(q)$ weakly depends on scale parameter q and whose long-range correlations and trends are easily destroyed, do not influence the structure of networks more than constant or randomized signal.

The properties of the time-varying signal do not influence the topological properties of small-world gel networks, Figure 3.5. Here model promotes the existence of hubs. As this is the mechanism through which the fluctuations alter the structure of evolving networks, the properties of the signal are not relevant.

3. Evolving complex network structure dependence on the properties of growth signals

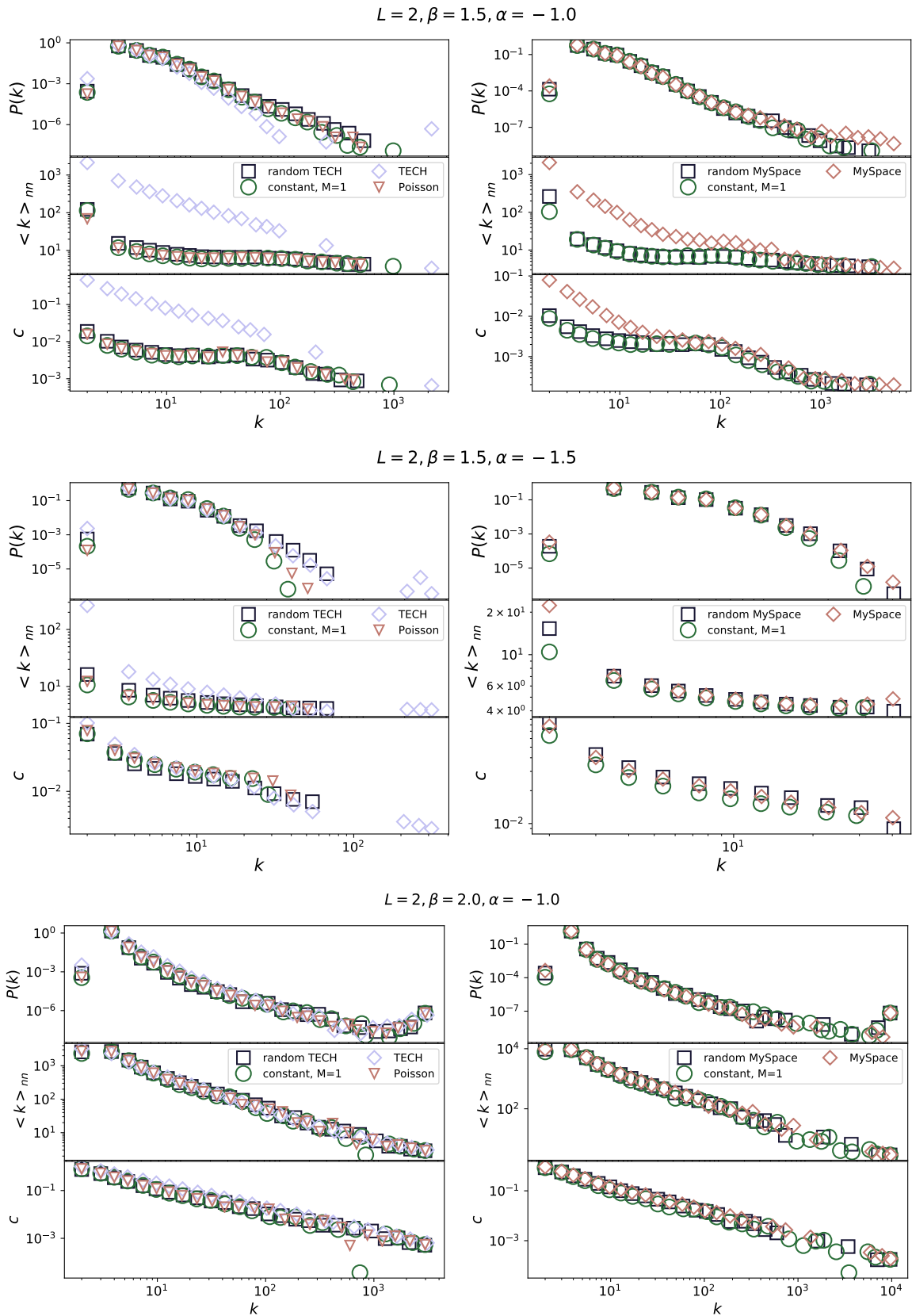


Figure 3.5: Degree distribution, the dependence of average first neighbor degree on node degree, the dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have value $\alpha = -1.0$, $\beta = 1.5$ and $L = 2$ for all networks. The networks are from the scale-free class. Model parameters have value $L = 2$, $\alpha = -1.5$, $\beta = 1.5$. The networks have stretched exponential degree distribution. Model parameters have value $L = 2$, $\alpha = -1.0$, $\beta = 2.0$. Generated networks have small-world properties.

3.2 Long range correlated signals

The previous section showed that the growth signal of real systems has complex dynamics. Besides long-range correlations, we also find multifractal properties, and it is hard to isolate individual effects and analyze their influence separately. When this is the case, synthetic signals with specific characteristics can help to verify our findings in real systems. The long-range correlated properties can be included in time series using Fourier filtering transform method [119].

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

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

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

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

The Fourier filtering method generates the Gaussian distributed data, so data are without broad distributions, nonlinear or multifractal properties. Using this method, we generated the signals for different values of the Hurst exponent; see Figure 3.6. The obtained signals are round to integers, and the mean values of signals are close to 4.

As before, we focus on the region of the model phase diagram with negative α and positive β as the transition line from stretched-exponential across scale-free to the small world-gel networks are found. We take a range of parameters $-3 \leq \alpha \leq -0.5$ and $1 \leq \beta \leq 3$ with steps 0.5, and we also vary the number of links each new node can create $L \in 1, 2, 3$. For each combination of (α, β, L) , we generate the sample of 100 networks and compare the structure of the network grown with fluctuating signals with different Hurst exponent $H \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$ and constant signal $M = 4$. The results represented by D-measure, shown in Figure 3.7, are obtained by averaging the D-measure between all possible pairs of generated networks.

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

3. Evolving complex network structure dependence on the properties of growth signals

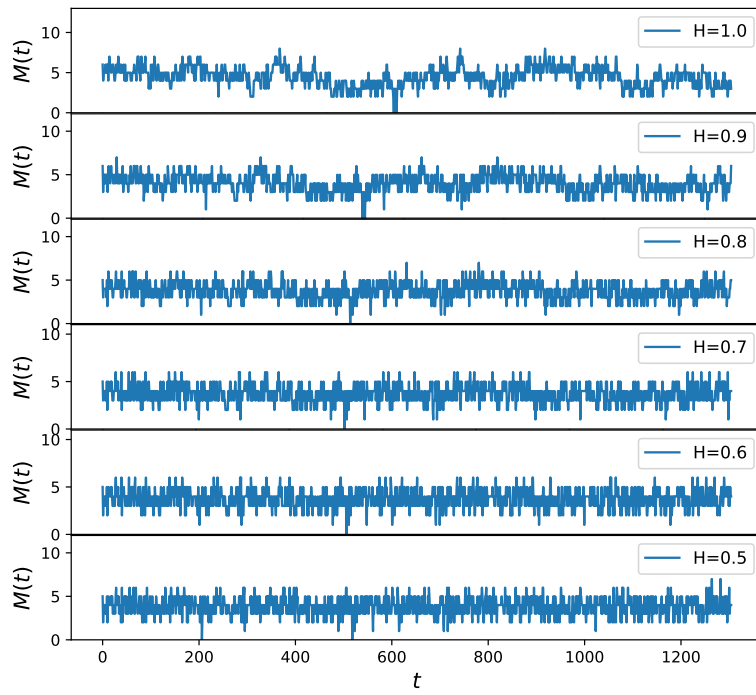


Figure 3.6: Monofractal signals generated with Fourier filtering method for different Hurst exponents

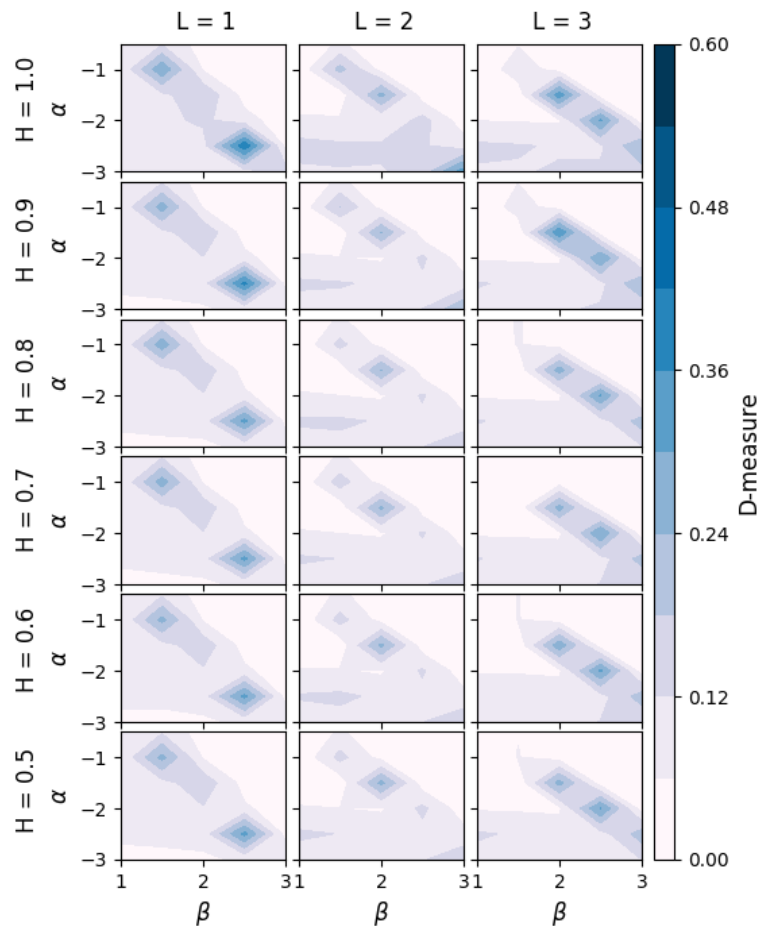


Figure 3.7: D-distance between networks generated with different long-range correlated signals with a fixed value of Hurst exponent and networks generated with constant signal $M=4$.

We further explore the assortativity index and clustering coefficient of generated networks. Figure 3.8 are results for several aging model parameters that show the difference between networks this model can produce. All networks are disassortative, with a negative degree-degree correlation index. For the parameters below critical line values, $\alpha = -2.5, \beta = 1.5$ r does not depend on the Hurst exponent. Above the critical line are small-world networks, and they are disassortative. The minimum value of the assortativity index is $r = -1$, for $L = 1$, indicating the presence of hubs connecting many nodes. The assortativity index grows slightly with link density.

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

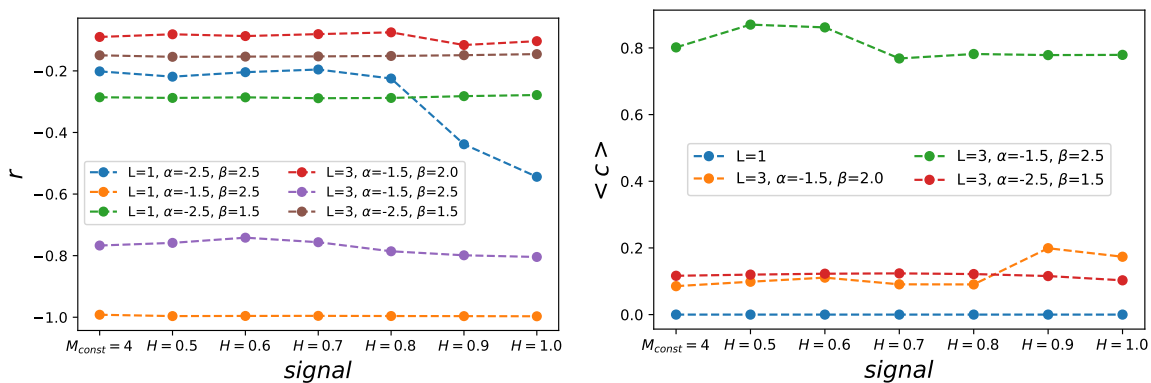


Figure 3.8: Mean assortativity index for networks generated with different model parameters α, β, L and different long-range correlated signals with Hurst exponent H .

Figure 3.8 shows the mean clustering coefficient. For $L = 1$, networks are uncorrelated trees with clustering coefficient 0. For network density $L > 1$, nodes are organized into clusters. Under the critical line, for the parameter, $L = 3, \alpha = -2.5, \beta = 1.5$, the clustering coefficient is constant and low. Similar values are obtained for the clustering coefficient for critical parameters $L = 3, \alpha = -1.5, \beta = 2.0$, but for Hurst exponent $H > 0.8$ clustering coefficient increases. Small world networks, $L = 3, \alpha = -1.5, \beta = 2.5$ are clustered, the value of $\langle c \rangle$ is high. The value of clustering for networks created with the constant signal is 0.8. Networks grown with white noise signals and signals with $H=0.6$ have higher clustering values, while networks grown with signals with a Hurst exponent larger than 0.6 have the same clustering value below 0.8.

3.3 Conclusions

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

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

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

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

Previous research on temporal networks [54] has shown that edge activation properties impact the complex system's dynamics. Also, different studies indicated the importance of fluctuating signals [27, 40, 35]. Our results imply that modeling the social and technological networks should include non-constant growth. In combination with local linking rules, the properties of growth signals can significantly alter the network structure.

Chapter 4

The growth of social groups

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

4.1 Empirical analysis of the social group growth

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

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

The Reddit data were downloaded from the <https://pushshift.io/> site. This site collects posts and comments daily; data are publicly available in JSON files for each month. The selected subreddits were created between 2006 and 2011, and we filtered those active in 2017. We removed subreddits active for less than two months. The obtained dataset has 17073 subreddits with 2195677 active members. For

¹<https://www.reddit.com/>

²www.meetup.com

each post, we extracted the subreddit-id, user-id, and the date when the user created the post. Finally, we selected the date when each user posted on each subreddit for the first time.

4.1.1 The empirical analysis of social groups

We have information about when the user attended the group event for each Meetup group. In contrast, we have detailed data about user activity for the subreddit, so we can extract the information when a user creates a post for the first time. Those dates are considered as the timestamp when a user joins to the group. So both datasets have the same structure: (g, u, t) , where t is the timestamp when user u joined group g . For each time step, we can calculate the number of new members in each group $N_i(t)$, and the group size $S_i(t)$. The group size at time step t is $S_i(t) = \sum_{k=t_0}^{k=t} N_i(k)$, where t_0 is month when group is created. The group size is increasing over time, as we do not have information if the user stopped to be active. Also, we calculate the growth rate as the logarithm of successive sizes $R = \log(S_i(t)/S_i(t-1))$.

Even though Meetup and Reddit are different online platforms, we find some common properties of these systems; see Figure 4.1. The number of groups and the number of new users grow exponentially. Still, subreddits are larger groups than Meetups. The distribution of groups sizes follows the lognormal distribution:

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

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

The distributions for Meetup group sizes in London and New York follow a similar lognormal distribution, with parameters $\mu = -0.93$, $\sigma = 1.38$ for London and $\mu = -0.99$ and $\sigma = 1.49$ for New York. The group sizes distribution of Subreddits is a broad lognormal distribution that resembles the power law; it has parameters $\mu = -5.41$ and $\sigma = 3.07$. Still, we used the log-likelihood ratio method and showed that lognormal distribution is a better fit for these data than the power-law. The Result section is given a detailed analysis that supports these findings.

The simplest model that generates the lognormal distribution is the multiplicative process [99]. Gibrat used this model to explain the growth of firms. The main assumption of this model is that growth rates $R = \log \frac{S_t}{S_{t-\Delta t}}$ do not depend on the size S and that they are uncorrelated. Further, this implies the lognormal distribution of the sizes, while the distribution of growth rates appears to be a normal distribution, [120], [121]. Figure 4.2 shows the distribution of the logrates that follow a lognormal distribution, contrary to the Gibrat law. Furthermore, logrates depend on the group size 4.2. For these reasons, the Gibrat law can not explain the growth of online social groups. Similar conclusions are shown in recent studies about cities or the growth of the internet [122, 123].

The growth of online social groups has universal behavior independent of the group's size. If we aggregate the groups created in the same year y , and each group size normalizes with average size $\langle S^y \rangle$, $s_i^y = S_i^y / \langle S^y \rangle$ we will find that group sizes distributions for the same dataset and different years fall on the same line, Figure 4.2. The same characteristics are observed for the distribution of the normalized logrates 4.2. The growth is universal over time, and the group sizes distribution does not change from year to year.

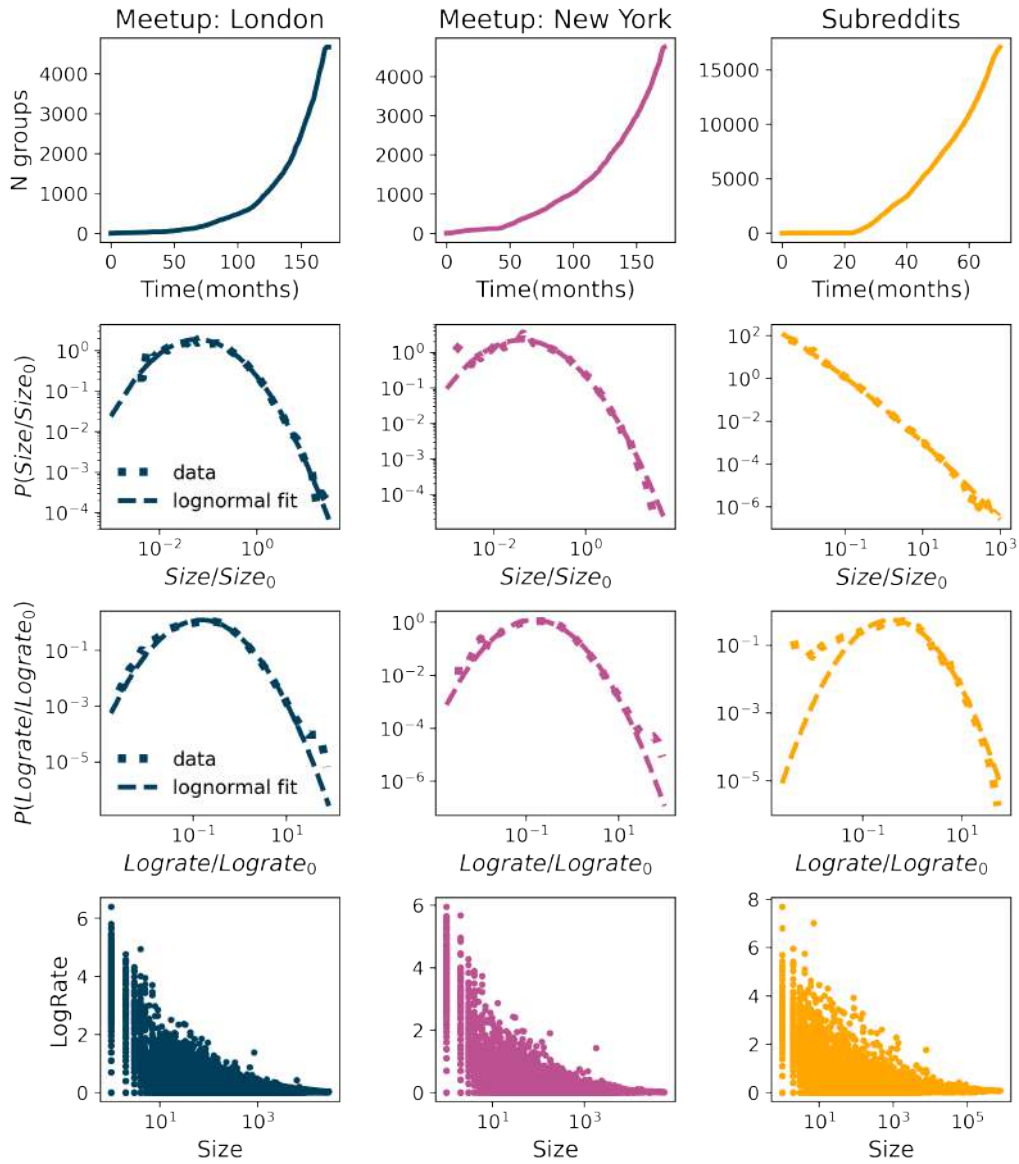


Figure 4.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011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

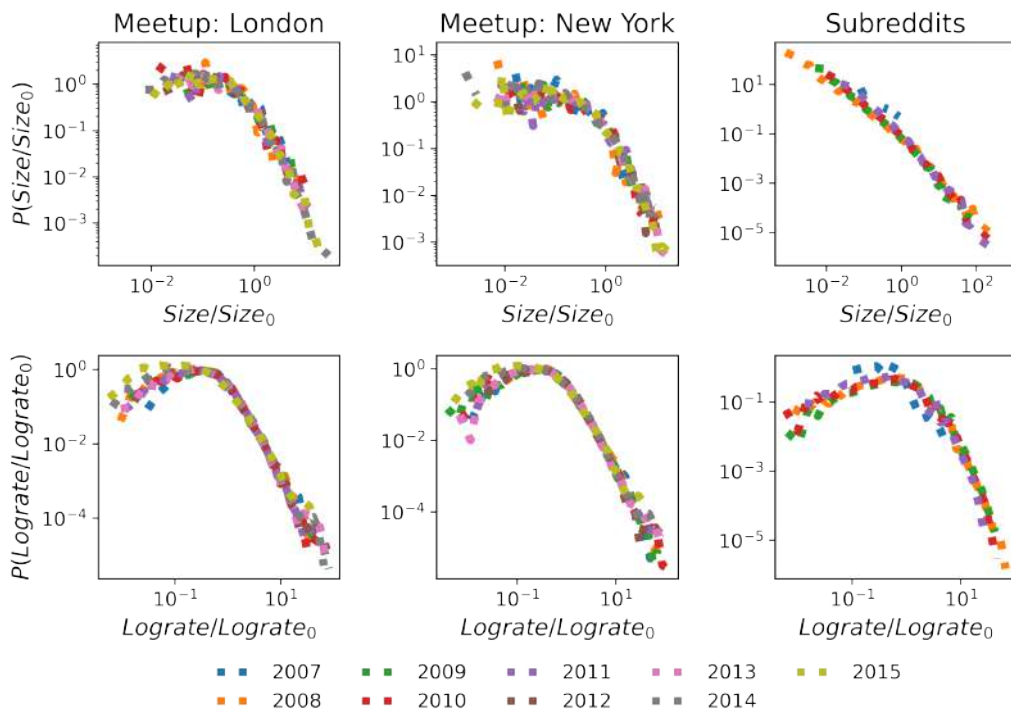


Figure 4.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

4.2 Theoretical model of social group growth

Meetup and Reddit engage members in different activities. Still, there are some underlying processes same in both systems. Each member can create new groups and join existing ones. Both systems grow in the number of groups and users, and each user can belong to an arbitrary number of groups. In the previous section, we identified the universal patterns in the growth of social groups, but the growth can not be modeled with the Gibrat law.

The complex network models allow us to simulate the growth of these systems considering all types of members' activities. We can identify how model parameters shape growth by varying linking rules. Regarding the user's group choice, it was shown that social connections play an important role [124, 125]. On the other hand, users can be driven by personal interests. Diffusion between groups could also be enhanced with rich-get-richer phenomena, where users join larger groups. With a complex network model, we can easily incorporate the nonlinear growth in the number of users and groups, as it is an important parameter that shapes the structure and dynamics of the complex network [126, 127, 63].

The evolution of the social groups has been studied using the co-evolution model in the reference [125]. This model consists of two evolving networks: the bipartite network, which stores connections between users and groups, and the affiliation network of social connections. At each time step, active users create new connections in the affiliation network; i.e., they make new friends. They also join existing groups or create new ones, which updates the bipartite network. The group selection can be random with probability proportional to the group size; otherwise, the group is selected through social contacts. Using this model, authors have reproduced the power-law group size distribution found in several communities, such as Flickr or LiveJournal. The empirical analysis of Meetup and Reddit groups showed that group size distribution could be lognormal, meaning that some different mechanisms control the growth of the groups.

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

4.2.1 Groups growth model

The representation of the model is given in Figure 4.3. The model consists of two networks:

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

The bipartite and social networks evolve. New users $N_U(t)$ are added to the network at each step. It is how the set of users V_U in the bipartite and social network can grow. At arrival, each new member connects to a randomly selected user in the social network G . This allows new members to choose a group based on social contacts [124]. The activity of old members is a stochastic process; old members

4. The growth of social groups

are activated with probability p_a . The set of active users \mathcal{A}_U has new members $N_U(t)$ and old members who decided to be active in that time step.

The active users can create a new group with probability p_g . By this, group node g is added to the set of group nodes V_G in bipartite network B . If an active user does not create a new group, it will join the existing one with probability $1 - p_g$, see lower panel on Figure 4.3. When the user creates a new group or joins an existing one, the link $e(u, g)$ is made in the bipartite network B .

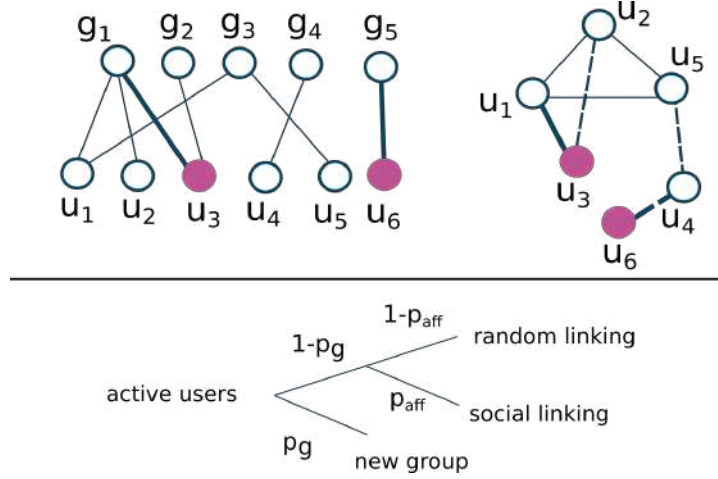


Figure 4.3: The top panel shows bipartite (member-group) and social (member-member) networks. Filled nodes are active members, while thick lines are new links in this time step. In the social network, dashed lines show that members are friends but do not share the same groups. The lower panel shows the model schema, where p_g is the probability that the user creates a new group, while p_{aff} is the probability that group choice depends on social connections. **Example:** member u_6 is a new member. First, it will make a random link with node u_4 , with probability, p_g makes a new group g_5 . With probability, p_a member u_3 is active, while others stay inactive for this time. Member u_3 will, with probability $1 - p_g$ choose to join one of the old groups, and with probability p_{aff} linking is chosen to be social. As its friend u_2 is a member of a group g_1 , member u_3 will also join group g_1 . When member u_3 joins group g_1 , it will make more social connections; in this case, it is member u_1 .

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

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

$$P_{ug} = \sum_{u_1 \in \mathcal{N}_g} A_{uu_1}. \quad (4.2)$$

After selecting group g , user u is introduced to new members in the group and can make new social contacts. In the simplest case, we could assume that all members belonging to a group are connected. However, previous research on this subject [117, 128, 125] has shown that the existing social connections of members in a social group are only a subset of all possible connections. We select X random members u_i from a group g and make new connections in the social network $e(u, u_i)$.

The model parameters p_a and p_g are important for controlling the number of users and groups. With larger parameter values p_a , more users become active, and the number of links in bipartite and social networks grows faster. Parameter p_g controls the rate at which new groups are created. For example, if $p_g = 0$, users will not create new groups. Also, if $p_g = 1$, users will only create new groups, and the resulting network will consist of star-like subgraphs. In real systems, we do not expect extreme values for probabilities p_a and p_g . First, not all members are constantly active, and we do not find a burst in the creation of the groups. From real data, we notice that there is always a higher number of users than groups in social systems. The parameter p_{aff} how users choose groups, and with higher p_{aff} social connections become more important.

4.2.2 Dependence of the group size distribution on model parameters

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

Figure 4.4 show the obtained group size distributions with power-law and lognormal fits. Users join randomly chosen groups for a lower parameter value $p_g = 0.1$ and $p_{aff} = 0$. Group size distributions are approximated with lognormal. When the affiliation parameter is larger, $p_{aff} = 0.5$, the lognormal distribution becomes broader, and so on, we find the larger maximum group size. If we increase the parameter $p_g = 0.5$, every second active user will create a group. At this group creation rate, the group size distribution deviates from lognormal, but it is not explained with power-law either, right column on Figure 4.4.

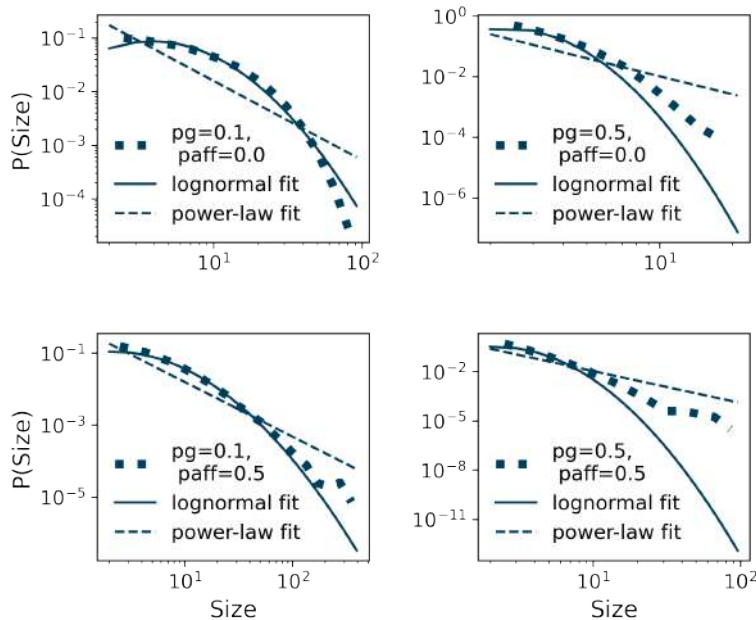


Figure 4.4: The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

Finally, we compare how group size distribution depends on different rules in random linking. In our model, the probability that the user chooses a random group is uniform. In contrast, in the co-evolution model [125], probability depends on the group size, as in the preferential attachment model. Instead of random linking, if we incorporate preferential linking, users with probability $1 - p_{aff}$ tend to choose larger groups, and group size distribution changes significantly. Similar to the co-evolution model, we find the power-law distribution. Figure 4.5 shows the results from a model where we add a constant number of new users at each time step. The probabilities p_a and p_g are fixed, and the affiliation parameter takes values 0, 0.5 and 0.8. If we consider random linking, a top panel on Figure 4.5, the distribution becomes broader with larger p_{aff} . On the other hand, with preferential linking, group size distribution is a power law, and the p_{aff} parameter does not have a large impact on the distribution shape.

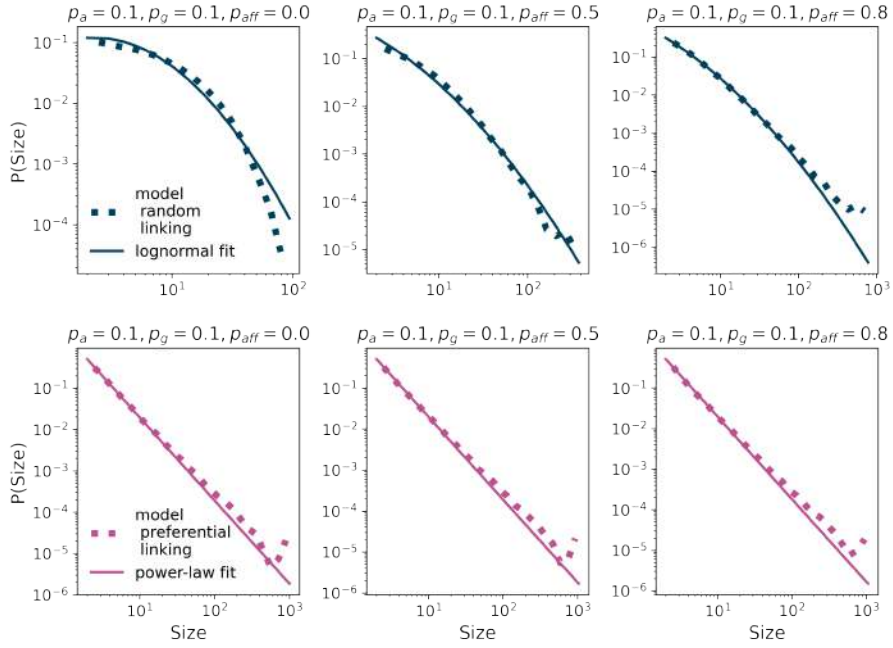


Figure 4.5: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff} . With probability,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

4.3 The growth of real social groups

The social systems do not grow at a constant rate. In Ref. [63], authors have shown that features of growth signal influence the structure of social networks. For these reasons, we use the real growth signal from Meetup groups located in London and New York and Reddit community to simulate the growth of the social groups in these systems. Figure 4.6 top panel shows the time series of the number of new members that join each of the three systems each month. All three systems have relatively low growth initially, which accelerates as the system becomes more popular.

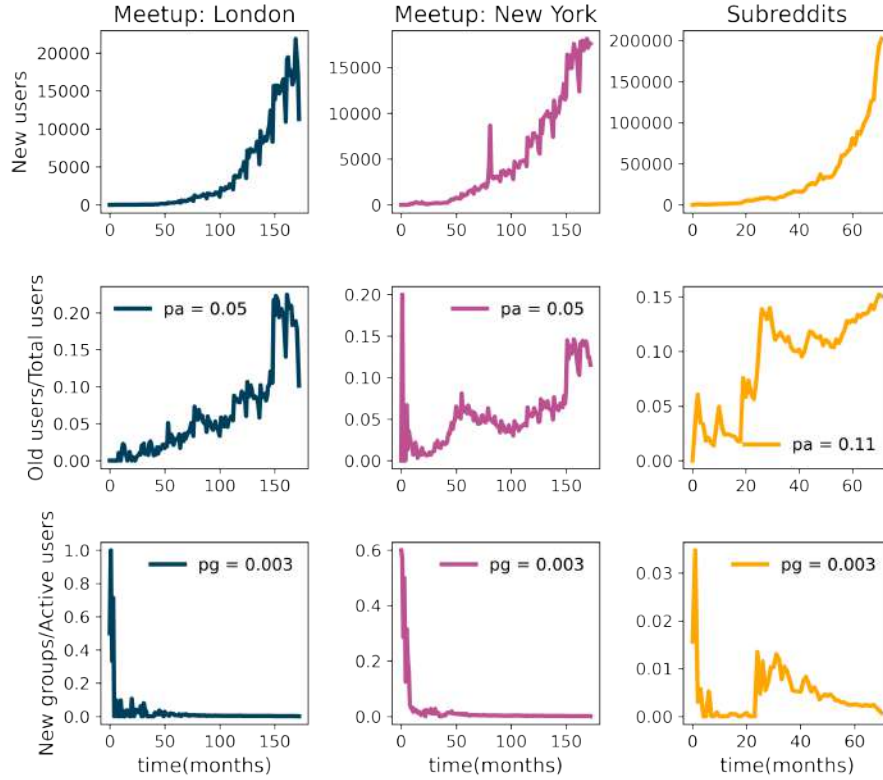


Figure 4.6: The time series of the number of new members (top panel), the ratio between old members and total members in the system (middle panel), and the ratio between new groups and active members (bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

We also use empirical data to estimate p_a , p_g and p_{aff} . Probabilities that old members are active p_a and that new groups are created p_g can be approximated directly from the data. Activity parameter p_a is the ratio between the number of old members active in month t and the total number of members in the system at time t . Figure 4.6 middle row shows the variation of parameter p_a during the considered time interval for each system. The values of this parameter fluctuate between 0 and 0.2 for London, and New York-based Meetup groups, while for Reddit, it ranges between 0 and 0.15. To simplify our simulations, we assume that p_a is constant in time and estimate its value as its median value during the 170 months for Meetup systems and 80 months of the Reddit system. For Meetup groups based in London and New York, $p_a = 0.05$, while Reddit members are more active on average, and $p_a = 0.11$ for this system.

Figure 4.6 bottom row shows the evolution of parameter p_g for the three considered systems. The p_g in month t is estimated as the ratio between the groups created in month t $Ng_{new}(t)$ and the total number of groups that month $Ng_{new}(t) + Ng_{old}(t)$, i.e., $p_g(t) = \frac{Ng_{new}(t)}{Ng_{new}(t) + Ng_{old}(t)}$. We see from Figure 4.6 that $p_g(t)$ has relatively high values at the beginning of the system's existence. In the beginning, these systems have a relatively small number of groups and often cannot meet the needs for the content of all

their members. As time passes, the number of groups and content offerings within the system grows, and members no longer have a high need to create new groups. Figure 4.6 shows that p_g fluctuates less after the first few months, and thus we again assume that p_g is constant in time and set its value to median value during 170 months for Meetup and 80 months for Reddit. For all three systems, p_g has the value of 0.003.

The affiliation parameter p_{aff} cannot estimate directly from the empirical data. For these reasons, we simulate the growth of social groups in each of the three systems with the time series of new members obtained from the real data and estimated values of parameters p_a and p_g , while we vary the value of p_{aff} . For each of the three systems, we compare the distribution of group sizes obtained from simulations for different values of p_{aff} with ones obtained from empirical analysis using Jensen Shannon (JS) divergence. The JS divergence [129] between two distributions P and Q is defined as

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

where $H(p)$ is Shannon entropy $H(p) = \sum_x p(x) \log(p(x))$. The JS divergence is symmetric, and if P is identical to Q , $JS = 0$. The smaller the JS divergence value, the better the match between empirical and simulated group size distributions. Table 4.1 shows the value of JS divergence for all three systems. We see that for London-based Meetup groups; the affiliation parameter is $p_{aff} = 0.5$, for New York groups $p_{aff} = 0.4$, while the affiliation parameter for Reddit $p_{aff} = 0.8$. Our results show that social diffusion is important in all three systems. However, Meetup members are more likely to join groups at random, while for Reddit members, their social connections are more important regarding the choice of the subreddit.

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

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

Figure 4.7 compares the empirical and simulation distribution of group sizes for three considered systems. We see that empirical distributions for Meetup groups based in London and New York are perfectly reproduced by the model and chosen values of parameters. In the case of Reddit, the distribution is very broad, and the model well reproduces the tail of the distribution. The bottom row of Figure 4.7 shows the distribution of logarithmic values of growth rates of groups obtained from empirical and simulated data. We see that the tails of empirical distributions for all three systems are well emulated by the ones obtained from the model. However, there are deviations that are the most likely consequence of using median values of parameters p_a , p_g , and p_{aff} .

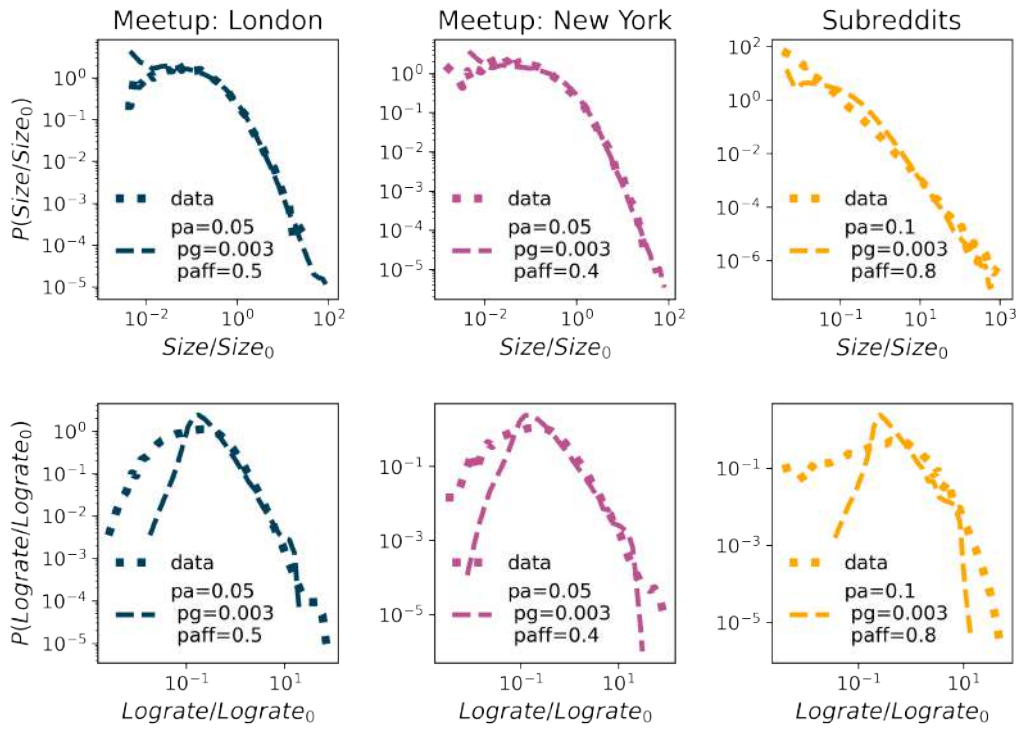


Figure 4.7: The comparison between empirical and simulation distribution for group sizes (top panel) and logrates (bottom panel).

4.3.1 Distributions fit

We compute the log-likelihood ratio R and p -value between different distributions and lognormal fit [103] to determine the best fit for the group size distributions. Distribution with a higher likelihood is a better fit. The log-likelihood ratio R has a positive or negative value, indicating which distribution represents a better fit. To choose between two distributions, we need to calculate the p -value to be sure that R is sufficiently positive or negative and that it is not the result of chance fluctuation from the result close to zero. If the p -value is small, $p < 0.1$, it is unlikely that the sign of R is the chance of fluctuations, and it is an accurate indicator of which model fits better.

Table 4.2 summarizes the findings for empirical data on group size distributions from Meetup groups in London and New York and Reddit. Using the maximum likelihood method, we obtain the parameters of the distributions [103]. The results indicate that lognormal distribution best fits all three systems. Figure 4.8 shows the distributions of empirical data and lognormal fit on data. For Meetup data, we present fit on stretched exponential distribution, which fits a large portion of data well. For subreddits, distribution is broad and potentially resembles power-law. Still, the lognormal distribution is a more suitable fit.

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

4. The growth of social groups

Table 4.2: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **groups sizes** of Meetup groups in London, New York and in Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

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

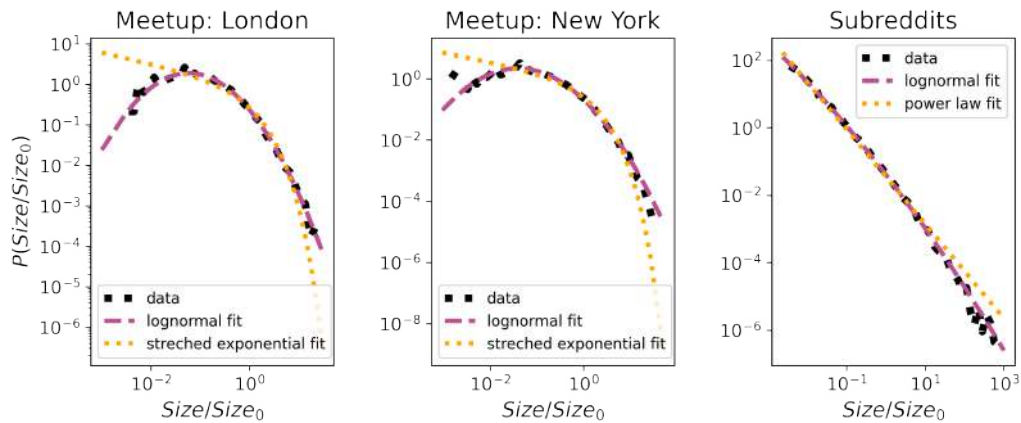


Figure 4.8: The comparison between lognormal and stretched exponential fit to London and NY data, and between lognormal and power law for Subreddits. The parameters for lognormal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits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

Table 4.3: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **simulated group sizes** of Meetup groups in London, New York and Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

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

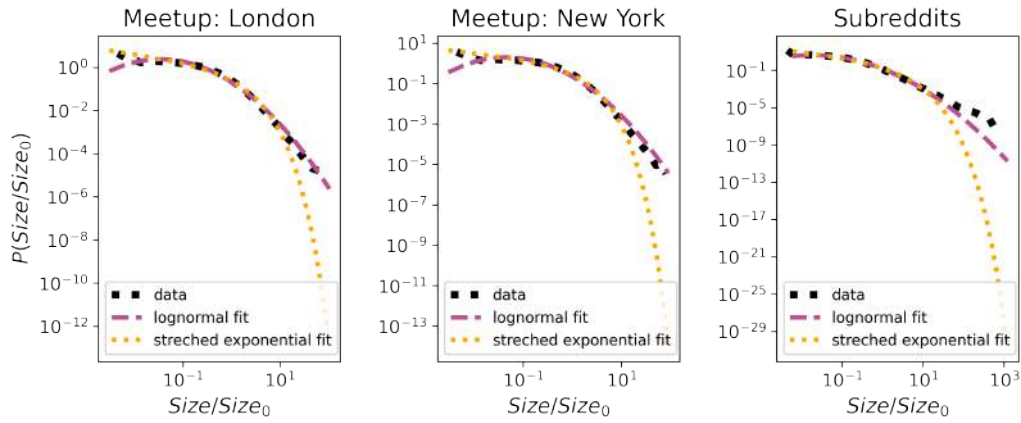


Figure 4.9: The comparison between lognormal and stretched exponential fit to simulated group size distributions. The parameters for lognormal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma =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

4.3.2 Users partition in bipartite network - degree distribution

So far, the group growth model has focused on the degree distribution of groups and under what rules the universalities in the system reflected in the lognormal distribution of group sizes emerge. The model parameter p_a controls the users' activity level; otherwise, it shapes the degree distribution of users in the bipartite network. As this probability is constant and uniform among all users, we do not expect rich properties of users' degree distribution. The expected distribution is exponential for growing random graph [130], and the groups' growth model produces the same property. In Figure 4.10, blue dots show degree distributions of modeled Meetup and Reddit systems. This distribution is very well fitted with exponential form. Furthermore, in empirical data, these distributions are long-tailed, green dots in Figure 4.10, so the model can not reproduce the degree distribution of the users. In real systems, the probability that the user is active does not have to be uniform and constant. The previous work proposed that each user has a specific lifetime [131], but different linking rules could play an important role in shaping users' degree distribution. For example, p_a could be preferential toward high-degree users or even be time-dependent.

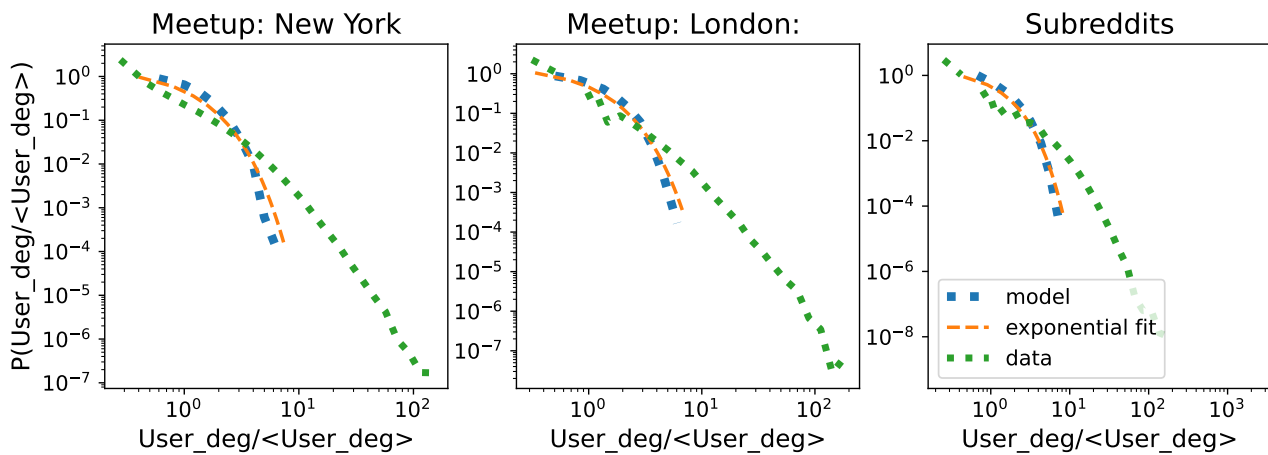


Figure 4.10: Users degree distributions from empirical data are compared to degree distributions observed by groups growth model.

4.4 Conclusions

We apply complex network theory and statistical physics methods to describe the evolution of online social groups, Meetups in London and New York and Reddits. Instead of studying user interaction networks in a single group, which is a common approach, we are interested in quantifying how users interact with the system of multiple groups and determining which processes drive the growth of groups. Similar systems have been analyzed before. For example, it was found that the distribution of the cities or firms follows the lognormal and stays stable, showing universal behavior. Contrary, the previous work on online social groups indicated that group size distributions of LiveJournal and Youtube follow power-law [125]. On the other hand, for Meetup and Reddit, we find the emergence of lognormal distribution of group sizes, and the distribution of Reddit is much broader. Furthermore, these systems grow exponentially in the number of groups and new users.

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

To explore the growth of these systems in more detail, we used a model where the social system is presented with evolving bipartite and social networks [125]. The bipartite network has partitions of users and groups, and a link exists if a user is a group member. The social network describes the social connections between members. At each time step, new users arrive in the system, following the time series of new users, and with probability, p_a old members also decide to be active. The active users can create a new group with probability p_g ; otherwise, they will join existing groups. Their decision to select a group based on social connection is determined with probability p_{aff} ; otherwise, the choice is random.

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

With results in this chapter, we contribute to the knowledge of the growth and segmentation of the socio-economic systems. Our work was motivated by the Co-evolution model [125]. The authors explore the social groups in which group size distribution scales as power-law. We identified different universality class, the system where group size distribution follows log normal. Further, we marked off a set of linking rules which led to lognormal group size distribution and compared these two cases. By this, we expanded the classes of social systems that can be modeled.

Chapter 5

The sustainability of evolving knowledge-based communities

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

The **Stack Exchange** (SE) is a network of question-answer websites on diverse topics. In the beginning, the focus was on computer programming questions with Stack Overflow¹ community. Its popularity led to the Stack Exchange network, which counts more than 100 communities on different topics. The SE communities are self-moderating, and the questions and answers can be voted, allowing users to earn Stack Exchange reputation and privileges on the site.

The new site topics are proposed through site Area51², and if the community finds them relevant, they are created. Every proposed StackExchange site needs interested users to commit to the community and contribute by posting questions, answers, and comments. After a successful private beta phase site reaches the public beta phase, other members can join the community. The site can be in the public beta phase for a long time until it meets specific SE evaluation criteria for graduation. Otherwise, it may be closed with a decline in users' activity. However, SE criteria for graduation have not been applied consistently on every SE site, as many sites graduated without reaching all required thresholds. As those measures only quantify the overall number of questions, answers, or highly active users, we want to understand how the SE community structure evolves and identify factors that influence sustainability. The need to share knowledge with others motivates users to use Q-A platforms. Still, the fact that they interact with each other reveals their sense of belonging to the community and the presence of trust among users. Our proxy for measuring trust in the community is the Dynamic Interaction Based Reputation Model.

¹More information about StackOverflow is available at <https://stackoverflow.co/>, and a broad introduction to the SE network is available at: <https://stackexchange.com/tour>.

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

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

5.1 Network properties of Stack Exchange data

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

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

The first approach is to aggregate all interactions in the first 180 days and study the properties of the static network. Many local and global network measures are dependent [12], and it was shown that degree distribution, degree-degree correlations, and clustering coefficient are sufficient for the description of the properties of complex networks [136].

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

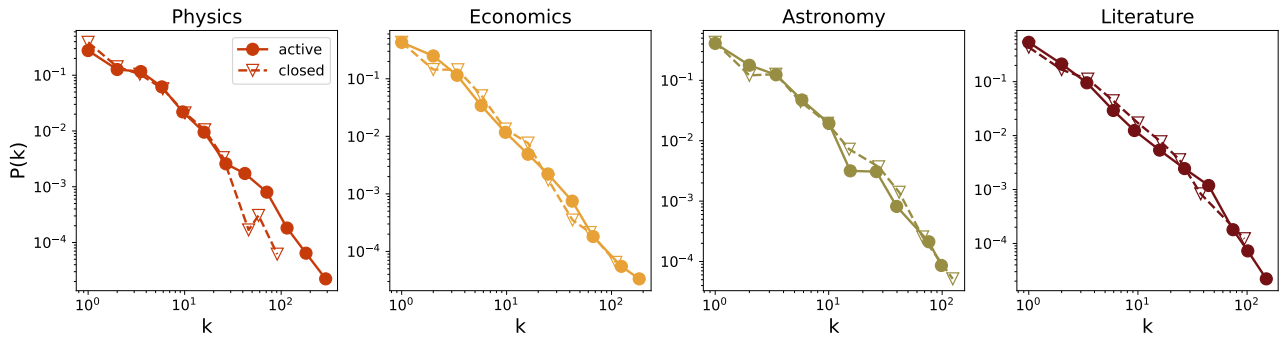


Figure 5.1: Degree distribution of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

If we take a look into **neighbor degree** depending on the node degree $k_{nn}(k)$, Figure 5.2, we find that there are structural differences between networks formed in the active and closed communities. On average, k -degree users in active communities have neighbors with a larger degree than is the case in closed communities. The results are consistent for Physics, Economics, and Literature. For Astronomy, we find different behavior, where the $k_{nn}(k)$ distributions of closed communities are on top of the distributions of the active ones.

A study on dynamics of social group growth shows that links between one's friends that are members of a social group increase the probability that that individual will join the social group [128].

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

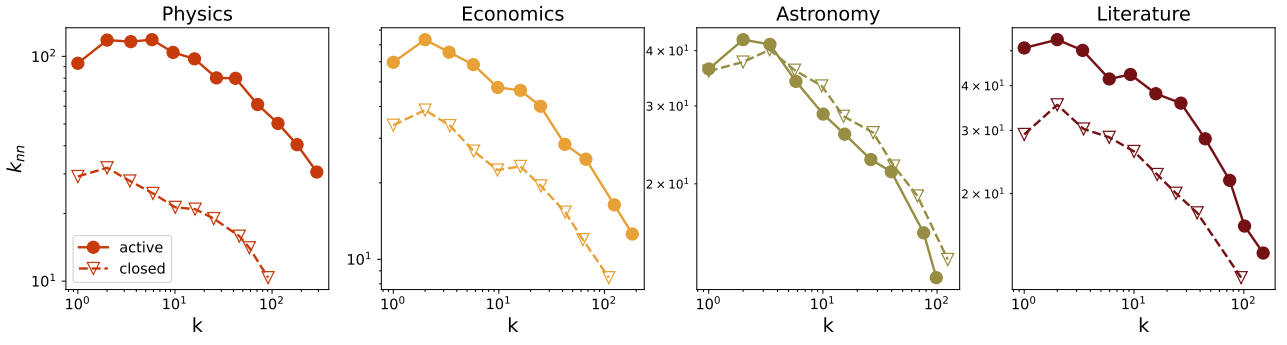


Figure 5.2: Neighbor degree dependence on the node degree of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

Furthermore, successful social diffusion typically occurs in networks with a high value of clustering coefficient [137]. These results suggest that high local cohesion should be a characteristic of sustainable communities. The dependence of the clustering coefficient on the node degree is shown in Figure 5.3. As expected, we find that active communities are more clustered.

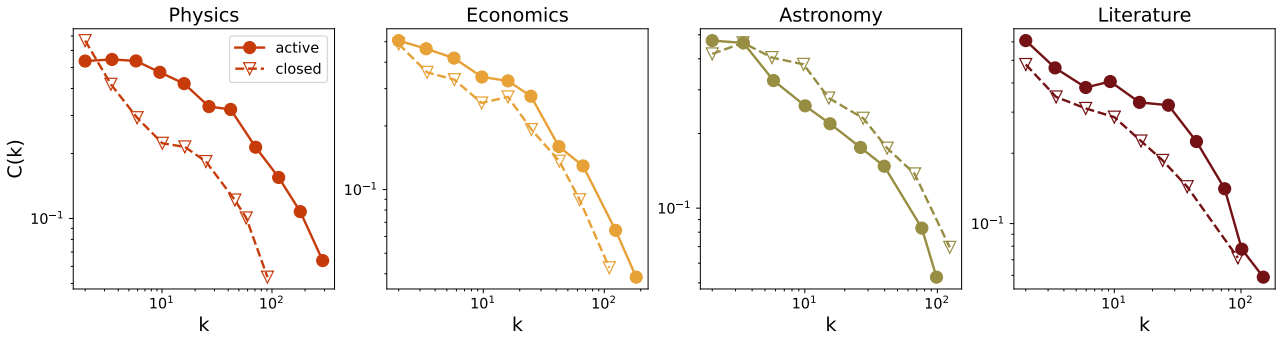


Figure 5.3: Clustering coefficient dependence on the node degree of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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

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

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

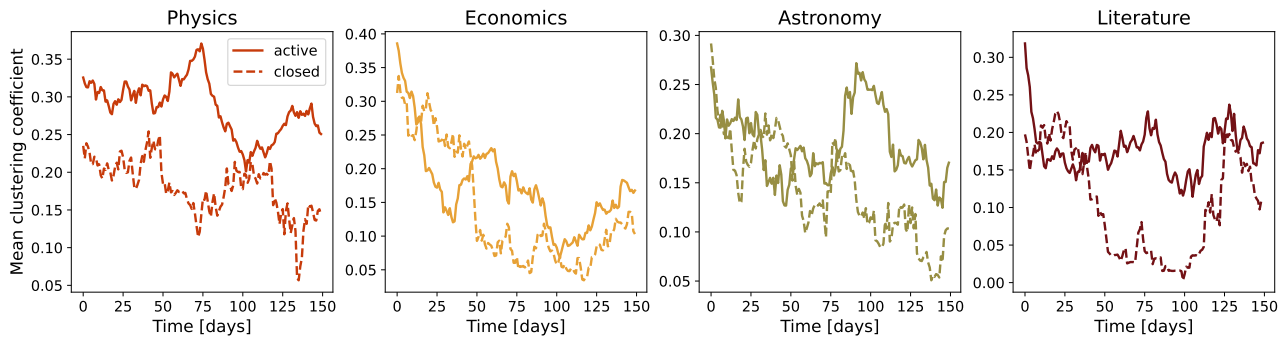


Figure 5.4: Mean clustering coefficient of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

phase, while later communities that are still active have higher clustering coefficient values. These results suggest that all communities have relatively high local cohesiveness and that lower clustering coefficient values may indicate its decline in the later phase of community life.

5.2 Core-periphery structure

Previous research on Stack Exchange communities has attempted to explain how different types of users interact. In Question-Answer communities are expected to be popular and casual users [138, 139]. Popular users generate the majority of interactions in the system; they are experts in the community and take care of answering questions and engaging the discussions through comments. As popular users, they considered the 10% of the most active users and showed that popular users are highly connected with themselves and casual users.

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

Although we find the difference between active and closed communities, the split according to 10% most active users does not guarantee that all popular users will be considered. Furthermore, the smaller group of frequently active users is similar to the core users in the core-periphery structure. This is why we will detect the core of each 30-day network. By this, separation is based on the network structure and is more consistent, as using the algorithmic approach, we optimize the connectivity inside the core, periphery, and among them. The core-periphery structure has a core that is a densely connected group of nodes, while the periphery has a low density [77, 66].

We use the Stochastic Block Model (SBM) to infer the core-periphery structure of each 30 days network snapshot and analyses how the core structure evolves. The SBM algorithm is adapted for inferring the core-periphery structure, [66]. For each 30 days network, we run the sample of 50 iterations and choose the model parameters according to the minimum description length. As stochastic models start from the random configuration, they can converge to different states, so we analyzed the stability of the inferred structures. More details are given in the appendix. We found that obtained structures differ, but the minimum description length does not fluctuate much. Also, different similarity measures between inferred core configurations take values higher than 0.9, indicating that the core structure is stable.

The Number of users in the core of active communities is higher than in closed communities, the

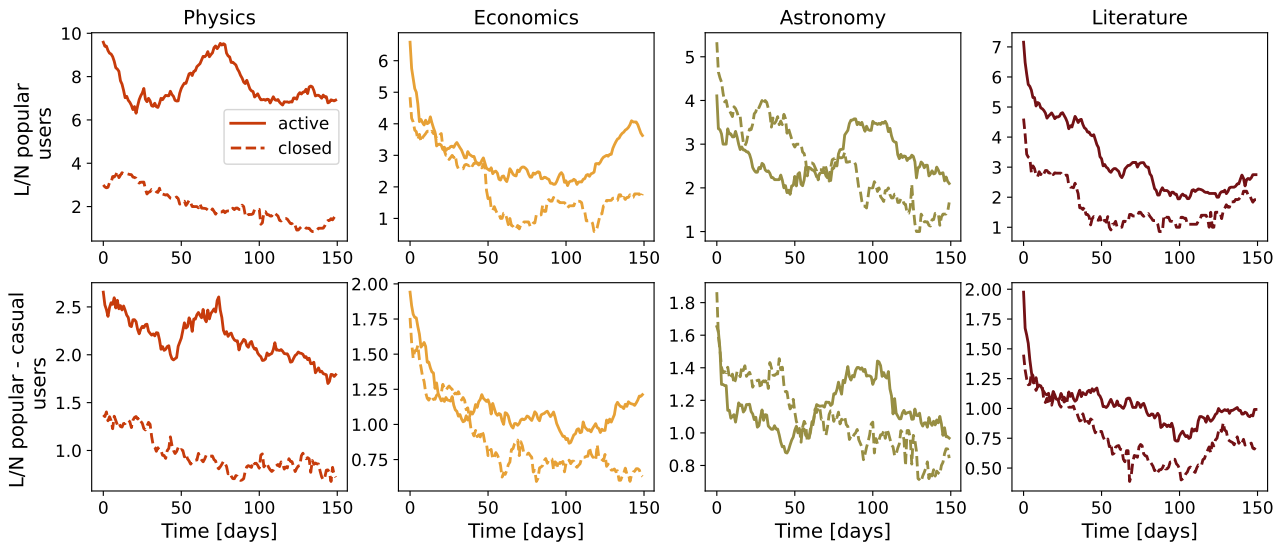


Figure 5.5: Links per node among popular users (top 10% of users) and between popular and casual users (everyone but popular users) of four pairs four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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

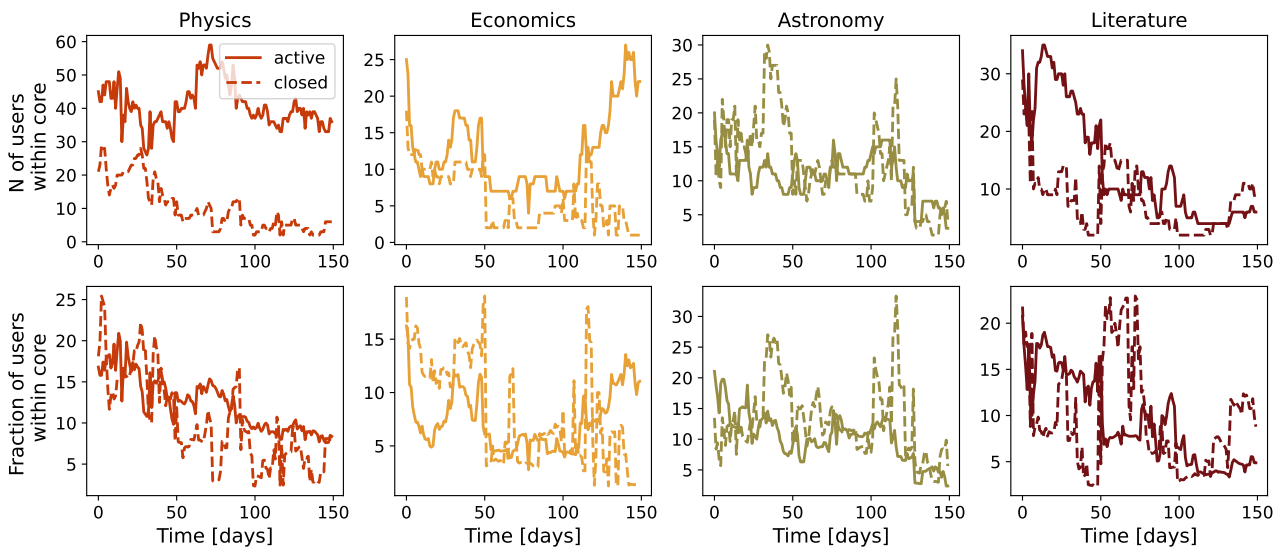


Figure 5.6: The size of the core (top) and a fraction of users in the core (bottom) of four pairs four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

The Number of users is constantly changing. To quantify the stability of the core structure, we compute the Jaccard's coefficient between core users in networks at time points t_1 and t_2 . The Jaccard coefficient range from 0 to 1, so the larger values of the Jaccard index indicate the more similar cores. The highest values are found around diagonal elements where we compare networks closer in time, see Figure 5.7. The core membership changes over time, and the change is more frequent in closed communities.

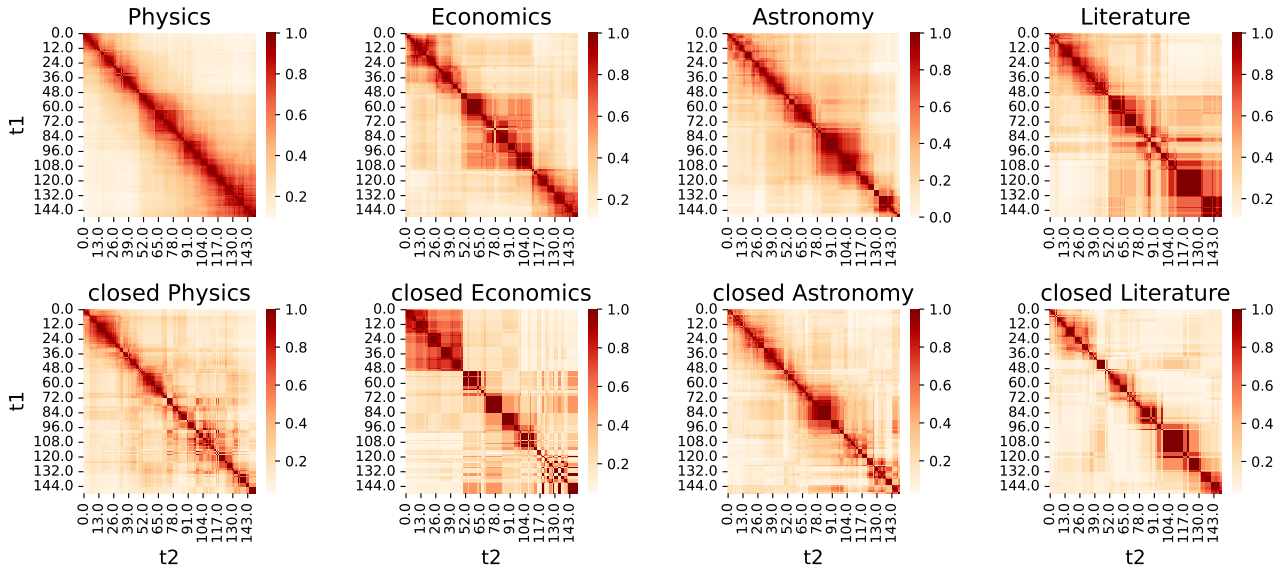


Figure 5.7: Jaccard index between core users in sub-networks at time points t_1 and t_2 for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

The average Jaccard index between cores in networks separated by time interval $t_i - t_j$ with the standard deviation confidence interval are shown in Figure 5.8. The Jaccard index decreases with the relative time difference between networks faster in closed communities. The relatively high overlap between distant networks confirms that active networks have a more stable core.

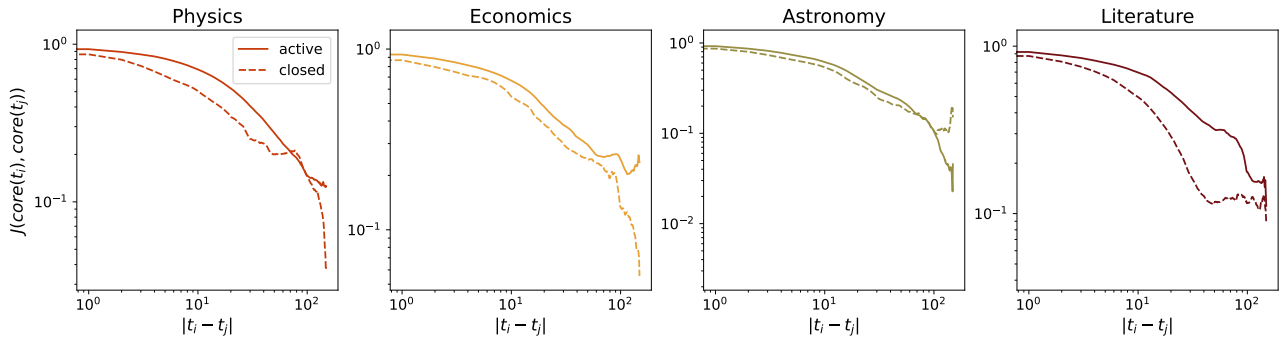


Figure 5.8: Jaccard index between core users in 30 days sub-networks for all possible pairs of 30 days sub-networks separated by time interval $|t_i - t_j|$ for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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

connectivity values in the closed site. Similar conclusions can be drawn for the connectivity between the core and periphery. The largest difference between active and closed sites is observed in Physics.

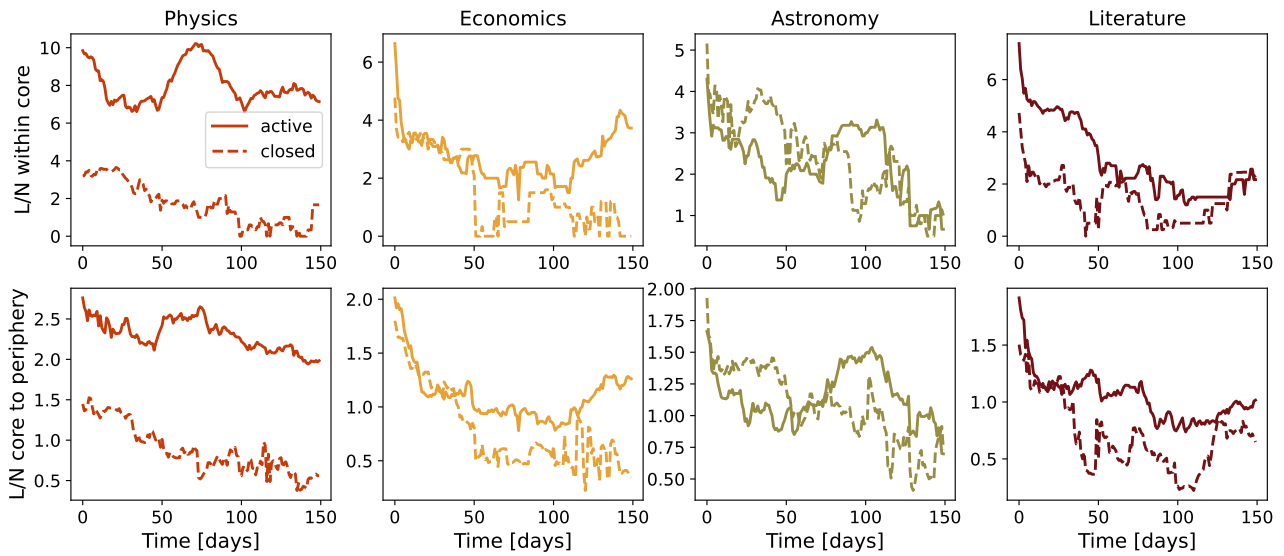


Figure 5.9: Number of links per node in core (top panel) and between core and periphery (bottom panel) for four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

5.3 Dynamical Reputation on Stack Exchange communities

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

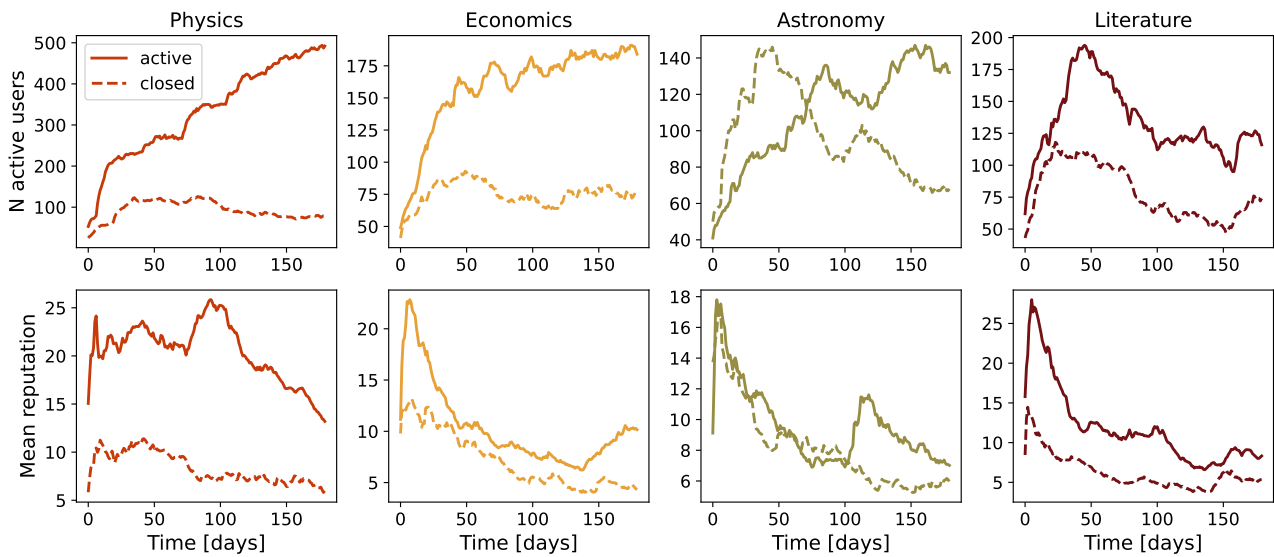


Figure 5.10: Number of active users (top panel) and mean reputation (bottom panel) of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

Dynamical reputation model, introduced in section 2.6, has three parameters. We explored different parameter combinations to find the set of parameters the most suitable for a given system of Stack Exchange communities. First, the basic reputation is set to $I_{bn} = 1$. The cumulative factor is $\alpha = 2$, as we want to emphasize the frequent interactions. The parameter β controls the reputation decay due to user inactivity. After the last activity, the user has a positive reputation for some period and is still impacting the other users. We optimized the Number of users with a reputation larger than 1 according to the number of users in the 30 days network and concluded that parameter $\beta = 0.96$. A more detailed discussion about the choice of parameters is in the appendix B.

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

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

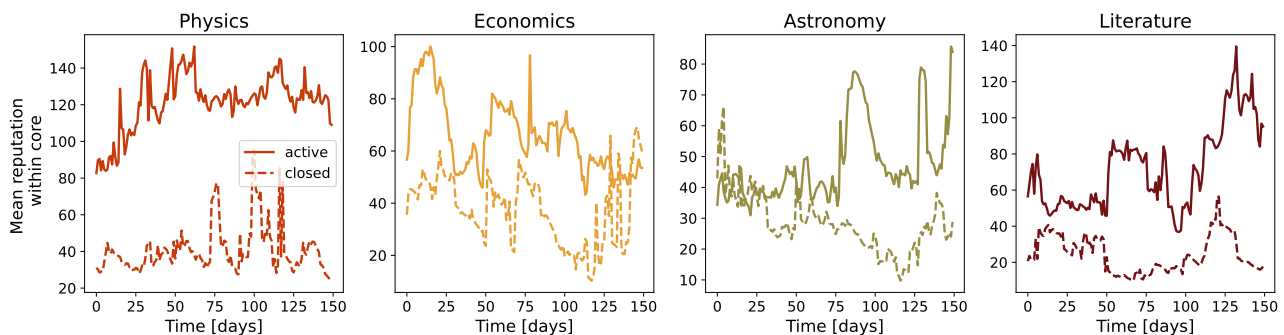


Figure 5.11: Dynamical reputation within the core of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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

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

Further, we investigate how the properties of user interaction networks correlate with the user's reputation. For example, we can measure the assortativity coefficient among connected users in the network. For each 30 days user interaction network, we calculate the reputation assortativity, using

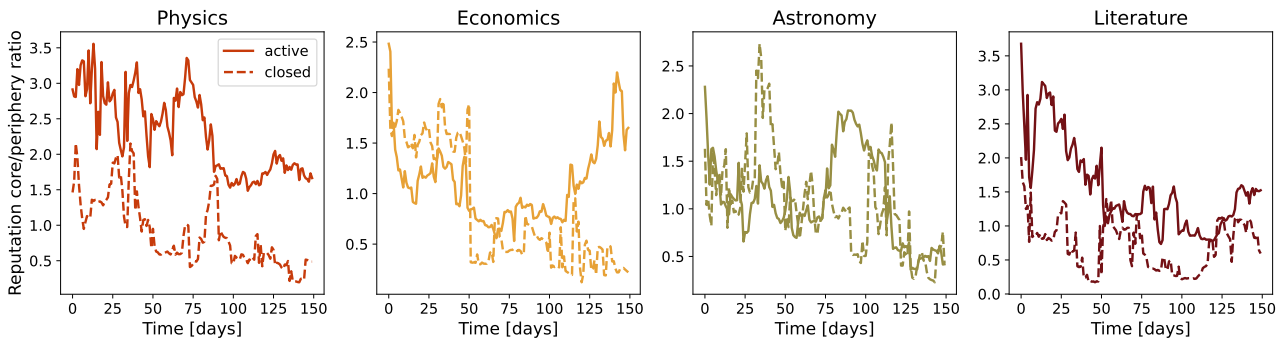


Figure 5.12: Ratio between the total reputation within network core and periphery of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

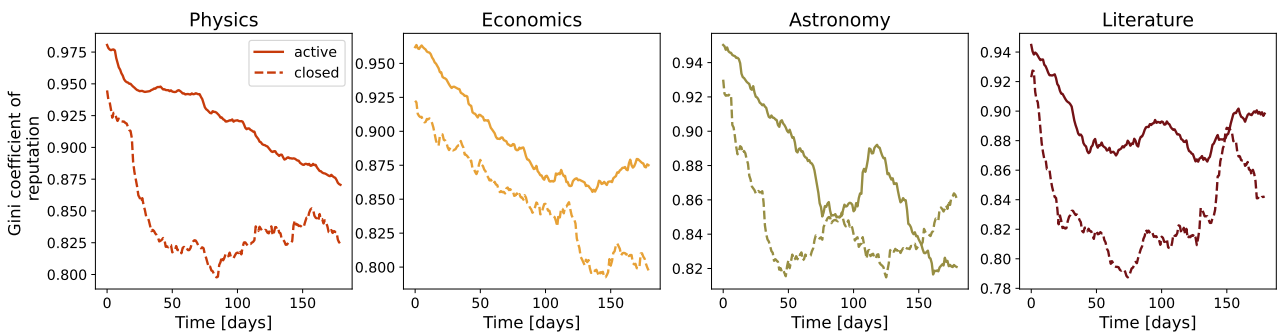


Figure 5.13: Gini index of dynamic reputation within the population of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

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

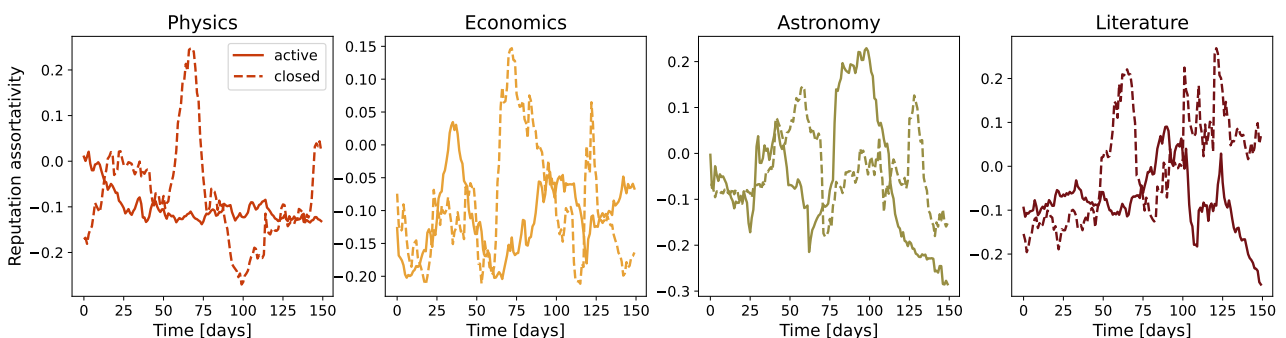


Figure 5.14: Dynamic Reputation assortativity in the network of interactions of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

Finally, we are interested in how dynamical reputation correlates with network measures. We compare the node's centrality in the 30-day network and the node's reputation on the last day of the 30-day sliding window. The correlation coefficient between dynamic reputation and node degree is very high; see the top panel on 5.15. The bottom panel shows correlations between dynamic reputation and betweenness centrality in the network, which are also high. We find that correlations are mostly higher in active communities; only for astronomy do they take similar values during the observed period.

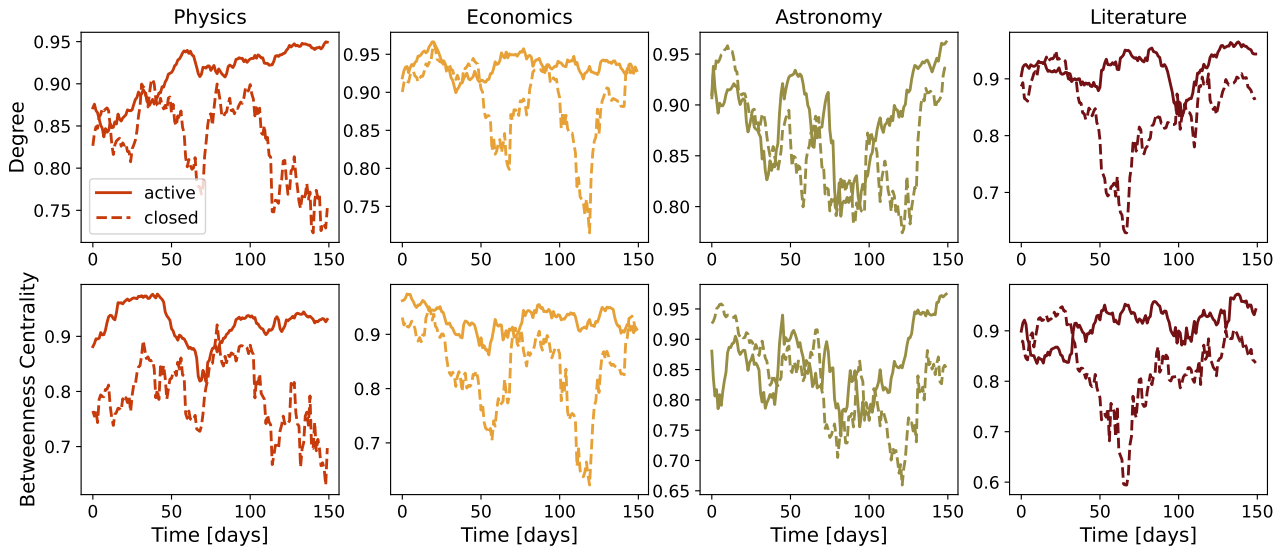


Figure 5.15: Coefficient of correlation between users' Dynamic Reputation and users' network degree (top) and users' betweenness centrality (bottom) of four pairs of Stack Exchange websites: Physics, Economics, Astronomy and Literature.

5.4 Conclusions

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

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

We studied how network properties evolve during the first 180 days. To closely examine the structure, we constructed the sub-networks within a 30days window. Sliding window by day, we continuously measure the structure of the network. The clustering coefficient is higher in active commu-

nities. The previous study suggested two groups of users in Q-A communities, popular and casual users [139]. This observation motivated us to closely analyze the network segmentation in the core-periphery structure. Based on Bayesian Stochastic modeling, we identify each 30-day network core user. Furthermore, using the DIBRM model[67], we quantify each user's reputation. This reputation is our proxy of trust, and its dynamics reflect some of the essential properties of trust. When a user is frequently active, the reputation increases; when inactivity declines, the user becomes less important.

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

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

Chapter 6

Conclusions

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

In chapter 3, we discussed how nonlinear growth signal shapes the structure of the complex network. The previous models combined linking rules with constant growth; however, empirical analysis of various real systems and agent-based simulation [39, 40] have indicated that properties of growth signal influence the dynamics of complex systems, as well as the structure of its interaction network. To investigate the connection between the features of the growth signal and the structure of an evolving network, we added one more parameter in the growth of the aging network model, the fluctuating growth signal, and examined how network properties change with the signal features. The most considerable influence is found on scale-free networks. Many interaction networks from social, technological, or biological systems have scale-free structures; they are correlated and clustered. These results suggested that it is important to study growing signals' properties. Signals from natural systems show trends and cycles and are characterized by long-range temporal correlations. The structure of the generated complex networks depends on the signal properties, and it is necessary to quantify these properties as they affect the network's topology differently. For example, the most significant difference between networks generated with fluctuating and constant signals is found for signals with multi-fractal properties. This difference is more negligible for monofractal signals or uncorrelated white noise. Fluctuating signals promote the creation of hubs in the network and shorten the paths between nodes.

Chapter 4 presented the results of the universal characteristics of the growth of online social groups—the growth of the system influence the structure of the interaction network. The distribution of the sizes of the complex systems usually follows some universal curve. In many cases, it is lognormal or power-law. The distribution of the dimensions of the city sizes could be explained with Zipf law [140]. The number of citations scales as lognormal distribution [21]. In this thesis, we empirically analyzed the growth of online social systems. They consist of groups whose growth is universal. The empirical analyses of Meetup groups and Reddits showed their group size distribution follows

universal lognormal distribution, stable over time. This research aimed to examine the structure and dynamics of the interaction network. We proposed the bipartite group model to gain a deeper understanding of the factors that affect the growth of social groups in a complex system. The growth in this model is driven by fluctuating signals, similar to the paper presented in chapter 3: we use a time series of new members from Meetup and Reddit. The number of groups also grows as each user can create a new one; otherwise, the user joins the old group, and different linking rules determine his decision. One option is that the user joins a group where she already has friends; it's determined with affiliating probability p_{aff} , while with probability $1 - p_{aff}$, the user chooses a random group. Group size distribution in this model is lognormal. The width of the lognormal distribution depends on the probability p_{aff} ; it becomes broader with a larger probability p_{aff} .

In chapter 5, we focused on the factors that influence the sustainability of evolving complex networks. Specifically, we investigated the sustainability of social groups on the Question-Answers platform Stack Exchange. Each site goes through several phases before being successful and launched. During that period, the site may be closed. We selected several topics in which sites for the first time were closed, but in the second attempt, they survived and are still active. We provide a detailed analysis of active and closed Stack Exchange sites, compare their properties and identify what is crucial for the community's survival. We map user interactions observed in 30 days onto complex networks. Further, we slide the window by one day and follow the evolution of the network.

According to the clustering properties of these networks, sustainable communities have a higher value of local cohesiveness. We use the Bayesian stochastic block modeling approach [66] to determine the core-periphery structure of these networks. We find that sustainable communities develop stable, better-connected cores. To analyze the evolution of collective trust in SE communities, we modify the Dynamic InteractionBased Reputation Model [67] (DIBR) model. We use the DIBR model to measure the user's reputation based on the frequency of their activity and its evolution during the first 180 days. The trust between core members of active communities develops early and is higher than in closed communities during the first 180 days. The early emergence of a stable, trustworthy core may be a crucial factor in determining a knowledge-sharing community's sustainability.

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

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

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

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

Appendix A

Stack Exchange

Stack Exchange data are public and regularly released. As closed communities were active between 180 and 210 days, we extracted only the first 180 days of data. Given that the first few months can be crucial for the further development of the community [141], we are interested in the early evolution of Stack Exchange sites.

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

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

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

Table A.1: Percentage of negatively voted interactions.

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

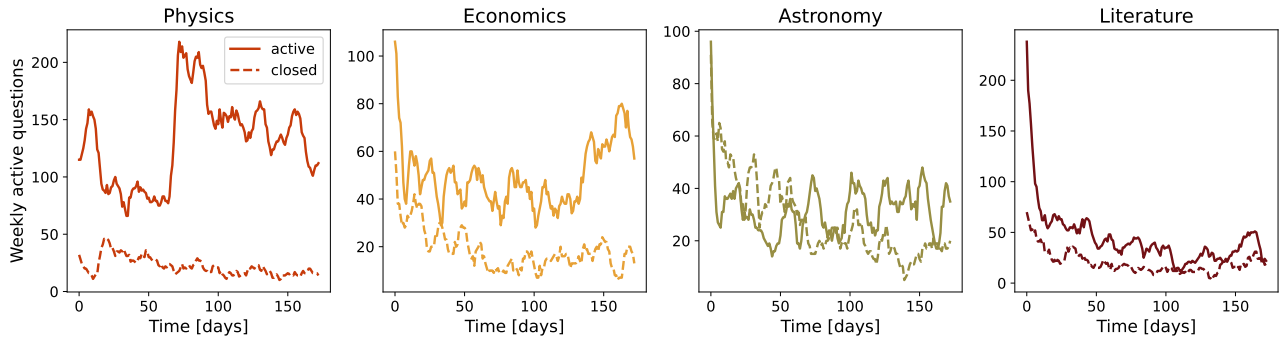


Figure A.1: Number of active questions within seven days sliding windows. Solid lines - active sites; dashed lines - closed sites.

A.1 Comparison between active and closed SE communities

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

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

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

Similarly, the official Stack Exchange community evaluation process considers simple metrics¹. To determine the success of sites they measure how many questions are answered, how many questions are posted per day, and how many answers are posted per question. There are two measures: the number of avid users and the number of visits that are not easily interpreted from the data. The site is *healthy* if it has ten questions per day, 2.5 answers per question and more than 90% of answered questions. For less than 80% of answered questions, five questions per day and 1 question per answer site *needs some work*.

We calculated Stack Exchange statistics for astronomy, economics, literature and physics and results are presented in Table A.3. After 180 days, only live physics is a healthy site while other live communities are at least in two criteria labelled as *okay*. Closed sites mostly *need some work*; the

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

exception is closed astronomy. For example, it has *excellent* percent of answered questions and *okay* answer ratio.

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

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

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

Appendix B

Selection of Dynamical Reputation Model parameters

The Dynamical Reputation Model(DIBRM) has several tuning parameters. In previous studies, the model [67, 142] was used to approximate real reputation on Stack Exchange sites [142], so model parameters were $t_a = 2, \beta = 1, \alpha = 1.4$, while the basic reputation value I_{bn} was +2 or +4. As $\beta = 1$, the forgetting factor is not considered. Our goal was to describe how reputation influences the sustainability of the community. Further, we wanted to resemble the concept of trust. Our tuning procedure differs from previous studies on Stack Exchange sites, and we ended up with different model parameters.

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

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

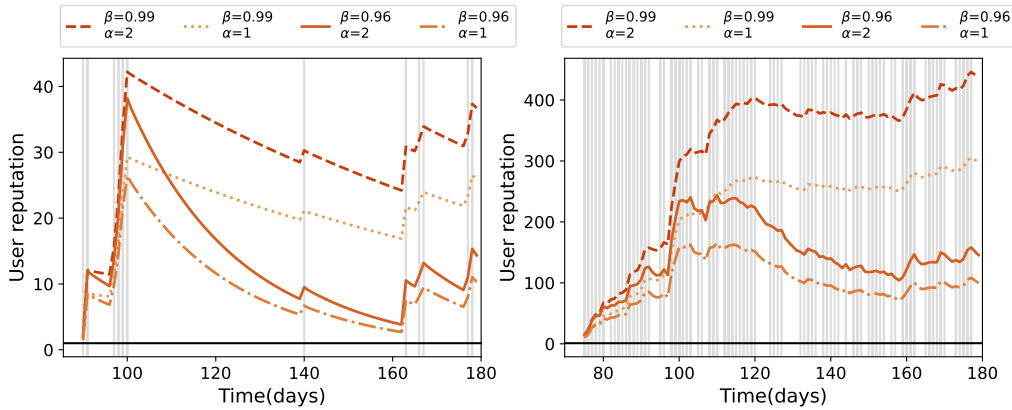


Figure B.1: Single users reputations, left panel shows sporadically active user, while user on right makes frequent interactions.

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

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

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

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

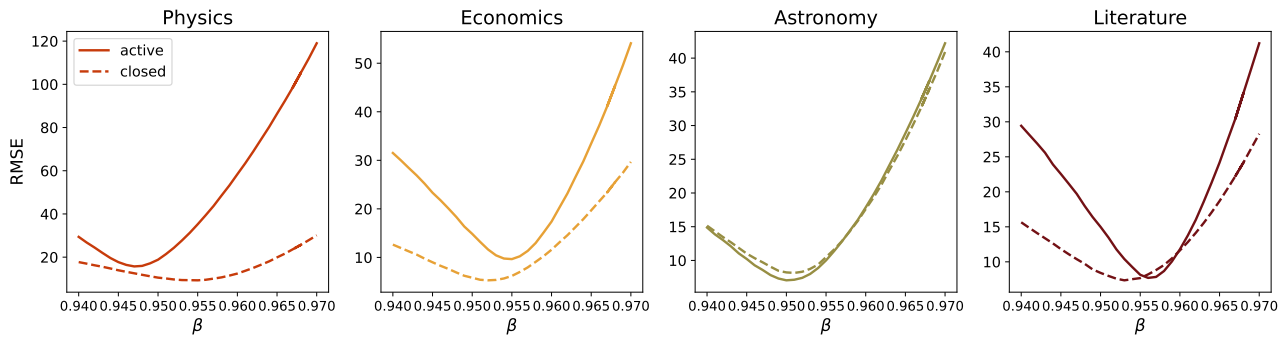


Figure B.2: RMSE between the number of active users in a sliding window of 30 days and the number of users with reputation > 1 for $0.94 < \beta < 0.97$ with step 0.001.

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

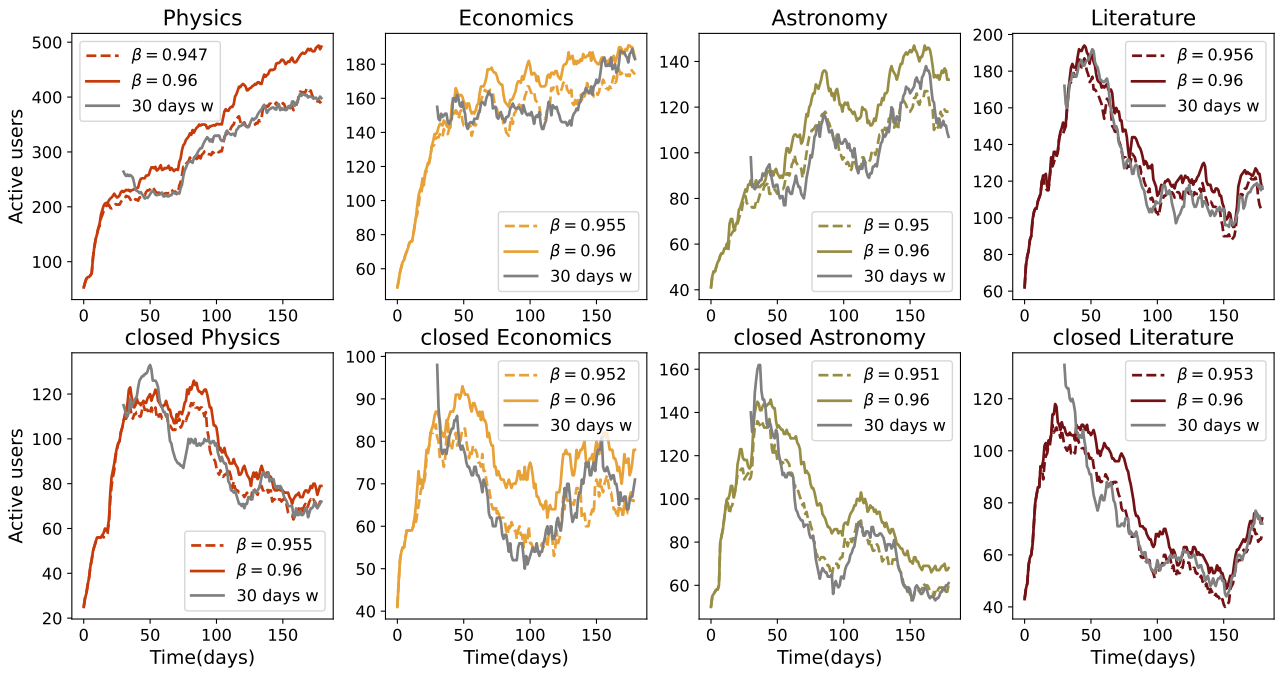


Figure B.3: Number of active users in a sliding window of 30 days and number of users with dynamic reputation higher than 1 for $\beta = 0.954$ and $\beta = 0.96$ which provide the best fit to the number of users in 30 days sub-networks for each community.

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

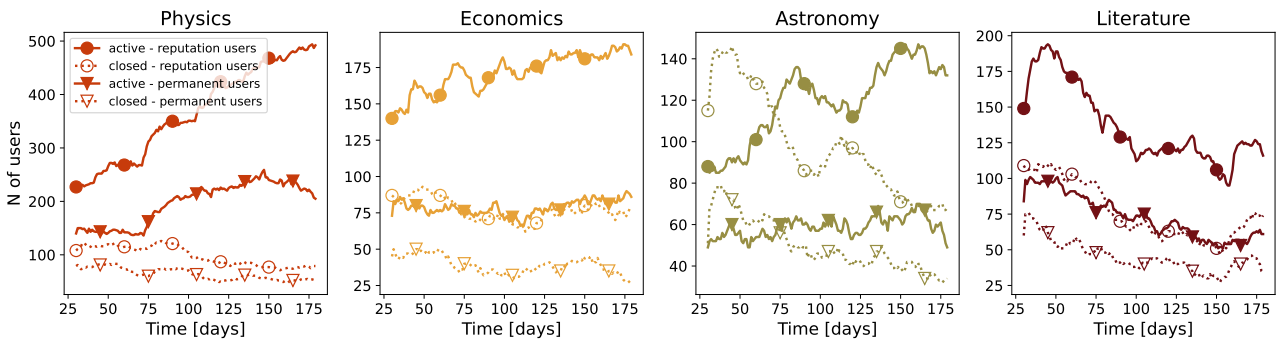


Figure B.4: Solid lines represent the number of users with dynamic reputation higher than 1 for $\beta = 0.96$ while dashed lines are the number of users within 30 days sliding window who were active and remained to be active.

Appendix C

The choice of the sliding window

To study the evolution of Stack Exchange communities, we chose to at each time step t analyze the structure of interaction networks created in the period $[t, t + \tau)$. By this, we have better insight into how network properties evolve. However, it is not defined what value the sliding window should take. The previous studies showed that the value of a sliding window determines how much information is saved. If τ is small, sub-networks are sparse, while for a large sliding window, important changes in the measures may not be detected [57, 58]. We analyze how network properties and dynamic reputation depend on the window size. For example, we use Astronomy and compare the active and closed communities, Figure C.1 Similar conclusions can be observed for other pairs of communities. The time window of 30 days approximates one month.

We show the network properties for sub-networks of 10, 30, and 60 days sliding windows. For a sliding window of 10 days, results may be too noisy, and we may not observe some important trends in the community. The number of users for beta astronomy seems to fluctuate around some mean value. On the larger scale, 30 days window, it is more apparent that the number of users slightly increases over time. Contrary, for too large an aggregation window (60 days), important information about the time series can be lost, such as the local minimum of the number of users around time step 80 that is observed for the 30-day sliding window. From network measures such as L/N and clustering, we conclude that the difference between closed and active sites is more transparent with a larger aggregation window. Still, on each scale, beta sites show a higher number of nodes, number of links per node and clustering coefficient.

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

C. The choice of the sliding window

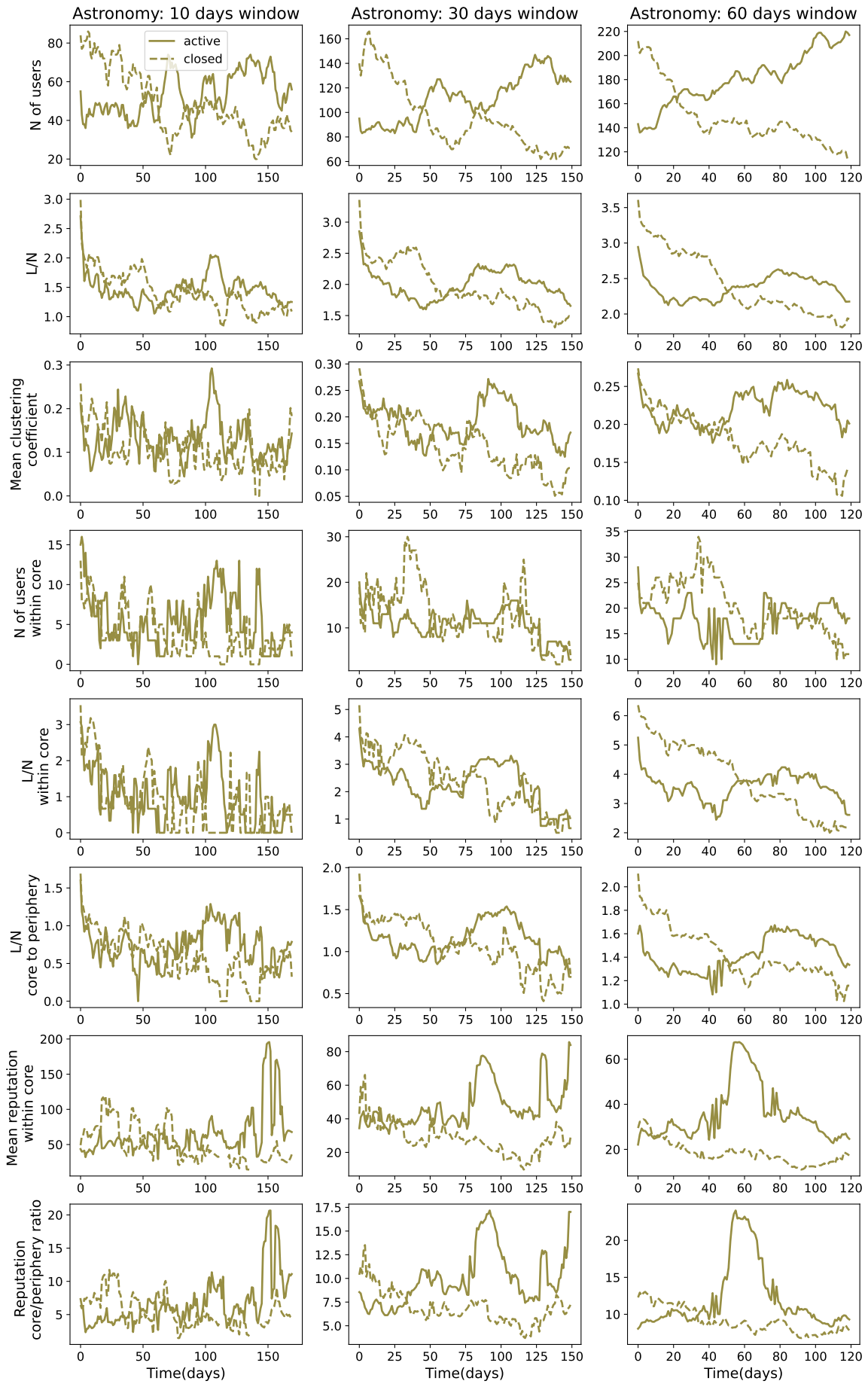


Figure C.1: Results for different sliding windows. For astronomy, solid blue lines- active, orange dashed lines - closed site.

Appendix D

Robustness of core-periphery algorithm

Precision and recall

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

The core-periphery structure has two groups, so confusion matrix [143] can be defined as:

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

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

From the confusion matrix, we can write the precision $P = n_{TP}/(n_{TP} + n_{FP})$ and recall $R = n_{TN}/(n_{TN} + n_{FN})$. These measures range from 0 to 1. The precision (recall) corresponds to the proportion of instances predicted to belong (not belong) to the considered class and which indeed do (do not) [143].

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

$$F_1 = 2 \frac{P \cdot R}{P + R} = \frac{2n_{TP}}{2n_{TP} + n_{FN} + n_{FP}}. \quad (\text{D.1})$$

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

The **Jaccard's** coefficient is the ratio of two classes' intersection to their union [143]. It can also be expressed in terms of a confusion matrix:

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

Normalized mutual information (NMI) is similarity measure between to partitions C and C' based on information theory [144]:

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

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

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

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

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

Adjusted rand index. For the set of nodes V , with n nodes, consider all possible combination of pairs (v_i, v_j) . We can select the number of the pairs where nodes belong to the same group in both partitions, C and C' , denoted as a . Similarly, as b , we can define the number of pairs whose nodes belong to different groups in partitions. Then, unadjusted rand index [145] is given as $RI = \frac{a+b}{\binom{n}{2}}$, where $\binom{n}{2}$ is number of all possible pairs. The RI between two randomly assigned partitions is not close to zero; for that reason, it is common to use the adjusted rand index [146], defined as:

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

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

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

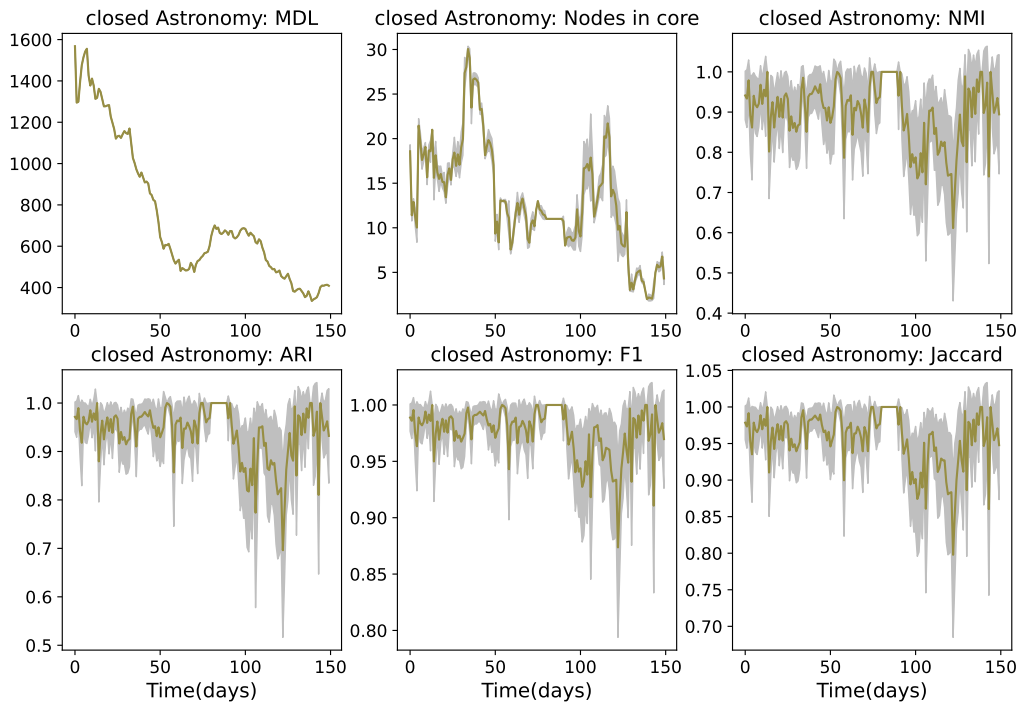


Figure D.1: Minimum description length, number of nodes in the core, normalized mutual information, adjusted rand index, F1 measure and Jaccard index, among 50 samples for 30-days sub-networks. Results are given for closed astronomy.

Bibliography

- [1] J. Kwapien and S. Drozd. Physical approach to complex systems. *Phys. Rep.*, 515:115–226, 2012.
- [2] Stefan Thurner, Rudolf Hanel, and Peter Klimek. 93Scaling. In *Introduction to the Theory of Complex Systems*. Oxford University Press, 09 2018.
- [3] S. A. Myers and J. Leskovec. The bursty dynamics of the twitter information network. In *Proceedings of the 23rd international conference on World wide web*, pages 913–924, 2014.
- [4] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer. Predicting scientific success based on coauthorship networks. *EPJ Data Science*, 3:9, 2014.
- [5] D. Fraiman, P. Balenzuela, J. Foss, and D. R. Chialvo. Ising-like dynamics in large-scale functional brain networks. *Phys. Rev. E*, 79:061922, 2009.
- [6] C. M. Schneider, L. de Arcangelis, and H. J. Herrmann. Modeling the topology of protein interaction networks. *Phys. Rev. E*, 84:016112, 2011.
- [7] L da F Costa, Francisco A Rodrigues, Gonzalo Travieso, and Paulino Ribeiro Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in physics*, 56(1):167–242, 2007.
- [8] Luciano da Fontoura Costa, Osvaldo N Oliveira Jr, Gonzalo Travieso, Francisco Aparecido Rodrigues, Paulino Ribeiro Villas Boas, Lucas Antiqueira, Matheus Palhares Viana, and Luis Enrique Correa Rocha. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. *Advances in Physics*, 60(3):329–412, 2011.
- [9] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45:167–256, 2003.
- [10] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013.
- [11] V. Latora, V. Nicosia, and G. Russo. *Complex networks: Principles, methods and applications*. 2017.
- [12] Stefano Boccaletti, Vito Latora, Yamir Moreno, Martin Chavez, and D-U Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4-5):175–308, 2006.
- [13] Parongama Sen and Bikas K Chakrabarti. *Sociophysics: an introduction*. Oxford University Press, 2014.

- [14] Frank Schweitzer. Sociophysics. *Phys. Today*, 71(2):40, 2018.
- [15] J. J. Binney, N. J. Dowrick, A. J. Fisher, and M. E. J. Newman. *The theory of critical phenomena: an introduction to the renormalization group*. Oxford University Press, 1992.
- [16] James P Sethna. *Statistical mechanics: entropy, order parameters, and complexity*, volume 14. Oxford University Press, USA, 2021.
- [17] Leo P Kadanoff. Scaling and universality in statistical physics. *Physica A: Statistical Mechanics and its Applications*, 163(1):1–14, 1990.
- [18] Antonios Garas, David Garcia, Marcin Skowron, and Frank Schweitzer. Emotional persistence in online chatting communities. *Scientific Reports*, 2(1):1–8, 2012.
- [19] Santo Fortunato and Claudio Castellano. Scaling and universality in proportional elections. *Physical review letters*, 99(13):138701, 2007.
- [20] Arnab Chatterjee, Marija Mitrović, and Santo Fortunato. Universality in voting behavior: an empirical analysis. *Scientific reports*, 3(1):1–9, 2013.
- [21] Filippo Radicchi, Santo Fortunato, and Claudio Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.
- [22] M. Barthelemy. The statistical physics of cities. *Nat. Rev. Phys*, 1:406–415, 2019.
- [23] Giorgio Fazio and Marco Modica. Pareto or log-normal? best fit and truncation in the distribution of all cities. *Journal of Regional Science*, 55(5):736–756, 2015.
- [24] Luís A Nunes Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, H Eugene Stanley, and Michael HR Stanley. Scaling behavior in economics: I. empirical results for company growth. *Journal de Physique I*, 7(4):621–633, 1997.
- [25] Michael HR Stanley, Luis AN Amaral, Sergey V Buldyrev, Shlomo Havlin, Heiko Leschhorn, Philipp Maass, Michael A Salinger, and H Eugene Stanley. Scaling behaviour in the growth of companies. *Nature*, 379(6568):804–806, 1996.
- [26] V. Verbavatz and M. Barthelemy. The growth equation of cities. *Nature*, 587:397–401, 2020.
- [27] Bernardo A Huberman and Lada A Adamic. Growth dynamics of the world-wide web. *Nature*, 401(6749):131–131, 1999.
- [28] S. Dorogovtsev. *Complex networks*. Oxford: Oxford University Press, 2010.
- [29] Albert-László Barabási. Scale-free networks: a decade and beyond. *science*, 325(5939):412–413, 2009.
- [30] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.
- [31] D. J. Watts and S. H. Strogatz. Collective dynamics of ‘small-world’ networks. *Nature*, 393:440–442, 1998.
- [32] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.
- [33] S. N. Dorogovtsev and J. F. F. Mendes. Evolution of networks with aging of sites. *Phy. Rev. E*, 62:1842, 2000.

- [34] Sergey N Dorogovtsev and José FF Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63(5):056125, 2001.
- [35] Jin Liu, Jian Li, Yadang Chen, Xianyi Chen, Zhili Zhou, Zejun Yang, and Cheng-Jun Zhang. Modeling complex networks with accelerating growth and aging effect. *Physics Letters A*, 383(13):1396–1400, 2019.
- [36] T. Pham, P. Sheridan, and H. Shimodaira. Joint estimation of preferential attachment and node fitness in growing complex networks. *Sci. Rep.*, 6:32558, 2016.
- [37] Parongama Sen. Accelerated growth in outgoing links in evolving networks: Deterministic versus stochastic picture. *Physical Review E*, 69(4):046107, 2004.
- [38] M. Mitrović and B. Tadić. Bloggers behavior and emergent communities in blog space. *The European Physical Journal B 2009 73:2*, 73(2):293–301, 2009.
- [39] Marija Mitrović and Bosiljka Tadić. Emergence and structure of cybercommunities. In *Springer Optimization and Its Applications*, volume 57, pages 209–227. Springer International Publishing, 2012.
- [40] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.
- [41] Reuven Cohen, Keren Erez, Daniel ben Avraham, and Shlomo Havlin. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.*, 85:4626–4628, Nov 2000.
- [42] Bosiljka Tadić. Dynamics of directed graphs: The world-wide web. *Physica A: Statistical Mechanics and its Applications*, 293(1-2):273–284, 2001.
- [43] Marija Mitrović and Bosiljka Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 80(2):026123, 2009.
- [44] Gourab Ghoshal, Liping Chi, and Albert-László Barabási. Uncovering the role of elementary processes in network evolution. *Scientific reports*, 3(1):1–8, 2013.
- [45] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al. Graph attention networks. *stat*, 1050(20):10–48550, 2017.
- [46] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.
- [47] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.
- [48] Chantat Eksombatchai, Pranav Jindal, Jerry Zitao Liu, Yuchen Liu, Rahul Sharma, Charles Sugnet, Mark Ulrich, and Jure Leskovec. Pixie: A system for recommending 3+ billion items to 200+ million users in real-time. In *Proceedings of the 2018 world wide web conference*, pages 1775–1784, 2018.
- [49] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M Bronstein. Fake news detection on social media using geometric deep learning. 2019. *arXiv preprint arXiv:1902.06673*, 1902.

- [50] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):1–10, 2019.
- [51] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.
- [52] Guido Caldarelli. *Scale-free networks: complex webs in nature and technology*. Oxford University Press, 2007.
- [53] Albert-László Barabási and Márton Pósfai. *Network science*. Cambridge University Press, Cambridge, 2016.
- [54] Petter Holme and Jari Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.
- [55] Naoki Masuda and Renaud Lambiotte. *A Guide to Temporal Networks*. 10 2016.
- [56] Petter Holme. Modern temporal network theory: a colloquium. *The European Physical Journal B*, 88(9):1–30, 2015.
- [57] Gautier Krings, Márton Karsai, Sebastian Bernhardsson, Vincent D Blondel, and Jari Saramäki. Effects of time window size and placement on the structure of an aggregated communication network. *EPJ Data Science*, 1(1):1–16, 2012.
- [58] Naomi A Arnold, Benjamin Steer, Imane Hafnaoui, Hugo A Parada G, Raul J Mondragon, Félix Cuadrado, and Richard G Clegg. Moving with the times: Investigating the alt-right network gab with temporal interaction graphs. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–17, 2021.
- [59] Mason A Porter. What is... a multilayer network. *Notices of the AMS*, 65(11), 2018.
- [60] Alberto Aleta and Yamir Moreno. Multilayer networks in a nutshell. *Annual Review of Condensed Matter Physics*, 10(1):45–62, 2019.
- [61] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. Multilayer networks. *Journal of Complex Networks*, 2(3):203–271, 07 2014.
- [62] Kamalika Basu Hajra and Parongama Sen. Phase transitions in an aging network. *Physical Review E*, 70(5):056103, 2004.
- [63] Ana Vranić and Marija Mitrović Dankulov. Growth signals determine the topology of evolving networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(1):013405, 2021.
- [64] Ana Vranić, Jelena Smiljanić, and Marija Mitrović Dankulov. Universal growth of social groups: empirical analysis and modeling. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(12):123402, 2022.
- [65] Ana Vranić, Aleksandar Tomašević, Aleksandra Alorić, and Marija Mitrović Dankulov. Sustainability of stack exchange q&a communities: the role of trust. *EPJ Data Science*, 12(1):4, 2023.
- [66] Ryan J Gallagher, Jean-Gabriel Young, and Brooke Foucault Welles. A clarified typology of core-periphery structure in networks. *Science advances*, 7(12):eabc9800, 2021.
- [67] A. Melnikov, J. Lee, V. Rivera, M. Mazzara, and L. Longo. Towards dynamic interaction-based reputation models. In *2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA)*, pages 422–428, 2018.

- [68] Ernesto Estrada and Philip A Knight. *A first course in network theory*. Oxford University Press, USA, 2015.
- [69] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *nature*, 406(6794):378–382, 2000.
- [70] Maarten Van Steen. Graph theory and complex networks. *An introduction*, 144, 2010.
- [71] Juyong Park and Mark EJ Newman. Origin of degree correlations in the internet and other networks. *Physical Review E*, 68(2):026112, 2003.
- [72] Mark EJ Newman. Assortative mixing in networks. *Physical review letters*, 89(20):208701, 2002.
- [73] Angélica Sousa da Mata. Complex networks: a mini-review. *Brazilian Journal of Physics*, 50:658–672, 2020.
- [74] Mark EJ Newman. Random graphs with clustering. *Physical review letters*, 103(5):058701, 2009.
- [75] Matthew O Jackson. *Social and economic networks*. Princeton university press, 2010.
- [76] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M Pardalos, Cristina Masoller, and Martín G Ravetti. Quantification of network structural dissimilarities. *Nature communications*, 8(1):1–10, 2017.
- [77] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3-5):75–174, 2010.
- [78] Martin Rosvall, Jean-Charles Delvenne, Michael T. Schaub, and Renaud Lambiotte. Different approaches to community detection. *CoRR*, abs/1712.06468, 2017.
- [79] Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics reports*, 659:1–44, 2016.
- [80] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science advances*, 3(5):e1602548, 2017.
- [81] Hocine Cherifi, Gergely Palla, Boleslaw K Szymanski, and Xiaoyan Lu. On community structure in complex networks: challenges and opportunities. *Applied Network Science*, 4(1):1–35, 2019.
- [82] Roger Guimera, Marta Sales-Pardo, and Luís A Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2):025101, 2004.
- [83] Clement Lee and Darren J Wilkinson. A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(1):1–50, 2019.
- [84] Tiago P Peixoto. Bayesian stochastic blockmodeling. *Advances in network clustering and block-modeling*, pages 289–332, 2019.
- [85] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
- [86] Benjamin H. Good, Yves-Alexandre de Montjoye, and Aaron Clauset. Performance of modularity maximization in practical contexts. *Phys. Rev. E*, 81:046106, Apr 2010.
- [87] A.-L. Barabási. Network science book. *Network Science*, 625, 2014.

- [88] Mark EJ Newman. Analysis of weighted networks. *Physical review E*, 70(5):056131, 2004.
- [89] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.
- [90] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences*, 104(1):36–41, 2007.
- [91] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical review E*, 74(1):016110, 2006.
- [92] Martin Rosvall, Jean-Charles Delvenne, Michael T Schaub, and Renaud Lambiotte. Different approaches to community detection. *Advances in network clustering and blockmodeling*, pages 105–119, 2019.
- [93] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical review E*, 83(1):016107, 2011.
- [94] Thorben Funke and Till Becker. Stochastic block models: A comparison of variants and inference methods. *PLOS ONE*, 14(4):1–40, 04 2019.
- [95] Peter Csermely, András London, Ling-Yun Wu, and Brian Uzzi. Structure and dynamics of core/periphery networks. *Journal of Complex Networks*, 1(2):93–123, 2013.
- [96] Fragkiskos D Malliaros, Christos Giatsidis, Apostolos N Papadopoulos, and Michalis Vazirgiannis. The core decomposition of networks: Theory, algorithms and applications. *The VLDB Journal*, 29(1):61–92, 2020.
- [97] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social networks*, 21(4):375–395, 2000.
- [98] Xiao Zhang, Travis Martin, and Mark EJ Newman. Identification of core-periphery structure in networks. *Physical Review E*, 91(3):032803, 2015.
- [99] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.
- [100] Mark EJ Newman. Power laws, pareto distributions and zipf’s law. *Contemporary physics*, 46(5):323–351, 2005.
- [101] Eckhard Limpert, Werner A Stahel, and Markus Abbt. Log-normal distributions across the sciences: keys and clues: on the charms of statistics, and how mechanical models resembling gambling machines offer a link to a handy way to characterize log-normal distributions, which can provide deeper insight into variability and probability—normal or log-normal: that is the question. *BioScience*, 51(5):341–352, 2001.
- [102] J. Nair, A. Wierman, and B. Zwart. *The Fundamentals of Heavy Tails*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2022.
- [103] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [104] Béla Bollobás and Oliver M Riordan. Mathematical results on scale-free random graphs. *Handbook of graphs and networks: from the genome to the internet*, pages 1–34, 2003.

- [105] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 74(1):47, 2002.
- [106] Paul L Krapivsky and Sidney Redner. Organization of growing random networks. *Physical Review E*, 63(6):066123, 2001.
- [107] Jan W Kantelhardt. Fractal and multifractal time series. *arXiv preprint arXiv:0804.0747*, 2008.
- [108] Chao Fan, Jin-Li Guo, and Yi-Long Zha. Fractal analysis on human dynamics of library loans. *Physica A: Statistical Mechanics and its Applications*, 391(24):6617–6625, 2012.
- [109] Sergei Sidorov, Alexey Faizliev, and Vladimir Balash. Fractality and multifractality analysis of news sentiments time series. *IAENG International Journal of Applied Mathematics*, 48(1), 2018.
- [110] Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American society of civil engineers*, 116(1):770–799, 1951.
- [111] Kun Hu, Plamen Ch Ivanov, Zhi Chen, Pedro Carpena, and H Eugene Stanley. Effect of trends on detrended fluctuation analysis. *Physical Review E*, 64(1):011114, 2001.
- [112] Jan W Kantelhardt, Eva Koscielny-Bunde, Henio HA Rego, Shlomo Havlin, and Armin Bunde. Detecting long-range correlations with detrended fluctuation analysis. *Physica A: Statistical Mechanics and its Applications*, 295(3-4):441–454, 2001.
- [113] Jan W Kantelhardt, Stephan A Zschiegner, Eva Koscielny-Bunde, Shlomo Havlin, Armin Bunde, and H Eugene Stanley. Multifractal detrended fluctuation analysis of nonstationary time series. *Physica A: Statistical Mechanics and its Applications*, 316(1-4):87–114, 2002.
- [114] E. Alexander F. E.A.F.I. Ihlen. Introduction to multifractal detrended fluctuation analysis in Matlab. *Front. Psychol.*, 3:141, 2012.
- [115] Albert-Laszlo Barabási, Hawoong Jeong, Zoltan Néda, Erzsebet Ravasz, Andras Schubert, and Tamas Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical mechanics and its applications*, 311(3-4):590–614, 2002.
- [116] Mark EJ Newman. The structure of scientific collaboration networks. *Proceedings of the national academy of sciences*, 98(2):404–409, 2001.
- [117] Jelena Smiljanić and Marija Mitrović Dankulov. Associative nature of event participation dynamics: A network theory approach. *PloS one*, 12(2):e0171565, 2017.
- [118] M. Šuvakov, M. Mitrović, V. Gligorijević, and B. Tadić. How the online social networks are used: dialogues-based structure of MySpace. *Journal of The Royal Society Interface*, 10:20120819, 2013.
- [119] Hernán A Makse, Shlomo Havlin, Moshe Schwartz, and H Eugene Stanley. Method for generating long-range correlations for large systems. *Physical Review E*, 53(5):5445, 1996.
- [120] Hernan Mondani, Petter Holme, and Fredrik Liljeros. Fat-tailed fluctuations in the size of organizations: the role of social influence. *PLoS One*, 9(7):e100527, 2014.
- [121] Dongfeng Fu, Fabio Pammolli, Sergey V Buldyrev, Massimo Riccaboni, Kaushik Matia, Kazuko Yamasaki, and H Eugene Stanley. The growth of business firms: Theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences*, 102(52):18801–18806, 2005.

- [122] Gerald F Frasco, Jie Sun, Hernán D Rozenfeld, and Daniel Ben-Avraham. Spatially distributed social complex networks. *Physical Review X*, 4(1):011008, 2014.
- [123] Jiang-Hai Qian, Qu Chen, Ding-Ding Han, Yu-Gang Ma, and Wen-Qing Shen. Origin of gibrat law in internet: Asymmetric distribution of the correlation. *Physical Review E*, 89(6):062808, 2014.
- [124] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 673–682, 2012.
- [125] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016, 2009.
- [126] Marija Mitrović, Georgios Paltoglou, and Bosiljka Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02005, 2011.
- [127] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.
- [128] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.
- [129] Jop Briët and Peter Harremoës. Properties of classical and quantum jensen-shannon divergence. *Phys. Rev. A*, 79:052311, May 2009.
- [130] Albert-László Barabási, Réka Albert, and Hawoong Jeong. Mean-field theory for scale-free random networks. *Physica A: Statistical Mechanics and its Applications*, 272(1-2):173–187, 1999.
- [131] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.
- [132] János Török and János Kertész. Cascading collapse of online social networks. *Scientific reports*, 7(1):1–8, 2017.
- [133] Thilo Gross and Bernd Blasius. Adaptive coevolutionary networks: a review. *Journal of the Royal Society Interface*, 5(20):259–271, 2008.
- [134] Xiao Han, Shinan Cao, Zhesi Shen, Boyu Zhang, Wen-Xu Wang, Ross Cressman, and H Eugene Stanley. Emergence of communities and diversity in social networks. *Proceedings of the National Academy of Sciences*, 114(11):2887–2891, 2017.
- [135] László Lőrincz, Júlia Koltai, Anna Fruzsina Győr, and Károly Takács. Collapse of an online social network: Burning social capital to create it? *Social Networks*, 57:43–53, 2019.
- [136] C. Orsini, M. Mitrović Dankulov, P. Colomer-de Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and Kriukov D. Quantifying randomness in real networks. *Nat. Commun.*, 6:8627, 2015.

-
- [137] Damon Centola, Víctor M Eguíluz, and Michael W Macy. Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374(1):449–456, 2007.
- [138] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Activity archetypes in question-and-answer (q&a) websites—a study of 50 stack exchange instances. *ACM Transactions on Social Computing*, 2(1):1–23, 2019.
- [139] Tiago Santos, Simon Walk, Roman Kern, Markus Strohmaier, and Denis Helic. Self-and cross-excitation in stack exchange question & answer communities. In *The World Wide Web Conference*, pages 1634–1645, 2019.
- [140] X. Gabaix. Zipf’s Law and the Growth of Cities. *Am. Econ. Rev.*, 89:129–132, 1999.
- [141] Yaniv Dover, Jacob Goldenberg, and Daniel Shapira. Sustainable online communities exhibit distinct hierarchical structures across scales of size. *Proceedings of the Royal Society A*, 476(2239):20190730, 2020.
- [142] Ekaterina Yashkina, Arseny Pinigin, JooYoung Lee, Manuel Mazzara, Akinlolu Solomon Adekotujo, Adam Zubair, and Luca Longo. Expressing trust with temporal frequency of user interaction in online communities. *Advances in Intelligent Systems and Computing*, pages 1133–1146, Cham, 2020. Springer International Publishing.
- [143] Vincent Labatut and Hocine Cherifi. Accuracy measures for the comparison of classifiers. *arXiv preprint arXiv:1207.3790*, 2012.
- [144] Leon Danon, Albert Diaz-Guilera, Jordi Duch, and Alex Arenas. Comparing community structure identification. *Journal of statistical mechanics: Theory and experiment*, 2005(09):P09008, 2005.
- [145] Jorge M Santos and Mark Embrechts. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer, 2009.
- [146] Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

Biography of the author

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

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

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

1. **Vranić A**, Tomašević A, Alorić A, Mitrović Dankulov M. Sustainability of Stack Exchange Q&A communities: the role of trust. *EPJ Data Science*. 2023 Feb 24;12(1):4.
2. **Vranić A**, Smiljanić J, Mitrović Dankulov M. Universal growth of social groups: empirical analysis and modeling. *Journal of Statistical Mechanics: Theory and Experiment*. 2022 Dec 7;2022(12):123402.
3. **Vranić A**, Mitrović Dankulov M. Growth signals determine the topology of evolving networks. *Journal of Statistical Mechanics: Theory and Experiment*. 2021 Jan 22;2021(1):013405.
4. **Vranić A**, Vučićević J, Kokalj J, Skolimowski J, Žitko R, Mravlje J, Tanasković D. Charge transport in the Hubbard model at high temperatures: Triangular versus square lattice. *Physical Review B*. 2020 Sep 21;102(11):115142.

Изјава о ауторству

Име и презиме аутора – **Ана Вранић**

Број индекса – **2017/8006**

Изјављујем

да је докторска дисертација под насловом

Evolving complex networks: structure and dynamics

(Растуће комплексне мреже: структура и динамика)

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

У Београду, 16.03.2023.

Потпис аутора

Ана Вранић

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора – **Ана Вранић**

Број индекса – **2017/8006**

Студијски програм – **Физика кондензоване материје и статистичка физика**

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

Растуће комплексне мреже: структура и динамика)

Ментор – **др Марија Митровић Данкулов**

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

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

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

У Београду, 16.03.2023.

Потпис аутора

Ана Вранић

Изјава о коришћењу

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

Evolving complex networks: structure and dynamics

(Растуће комплексне мреже: структура и динамика)

која је моје ауторско дело.

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

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

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

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

У Београду, 16.03.2023.

Потпис аутора

Ана Брашић

1. **Ауторство.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.
2. **Ауторство – некомерцијално.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.
3. **Ауторство – некомерцијално – без прерада.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.
4. **Ауторство – некомерцијално – делити под истим условима.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.
5. **Ауторство – без прерада.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.
6. **Ауторство – делити под истим условима.** Дозвољаваате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.