Vranić et al. *EPJ Data Science* (2023) 12:4
https://doi.org/10.1140/epjds/s13688-023-00381-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

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

Ana Vranić[1*], Aleksandar Tomašević[2], Aleksandra Alorić[1,3] and Marija Mitrović Dankulov[1]

*Correspondence: anav@ipb.ac.rs
[1]Institute of Physics Belgrade, University of Belgrade, Pregrevica 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.

Springer

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

[1] More 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.

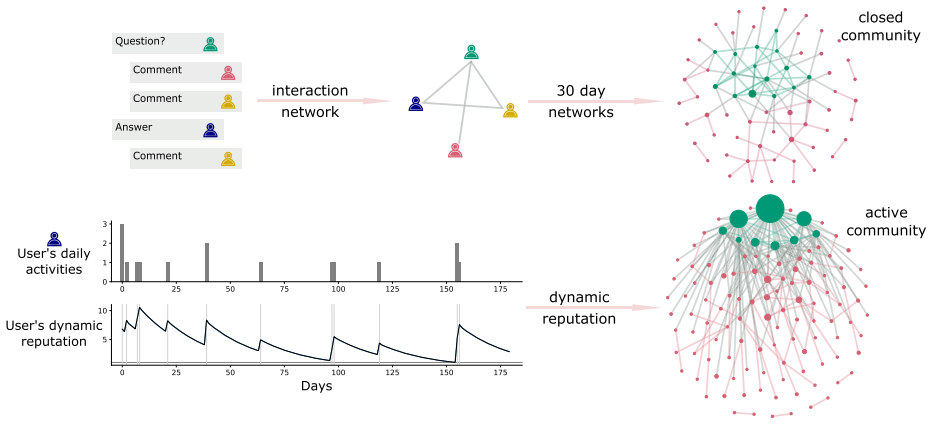

Figure 1 Visual abstract: Top row illustrates how user interaction via questions, answers, and comments is translated into an undirected network of interactions between users and finally aggregated over 30 day windows. The bottom row shows activity and corresponding dynamic reputation for one user from the closed Literature SE community. Networks on the right illustrate differences between closed and active SE Literature communities. Nodes are colored according to the core/periphery affiliation, while their size corresponds to dynamic reputation on the last day of interaction that the network contains

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

[2]For 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/

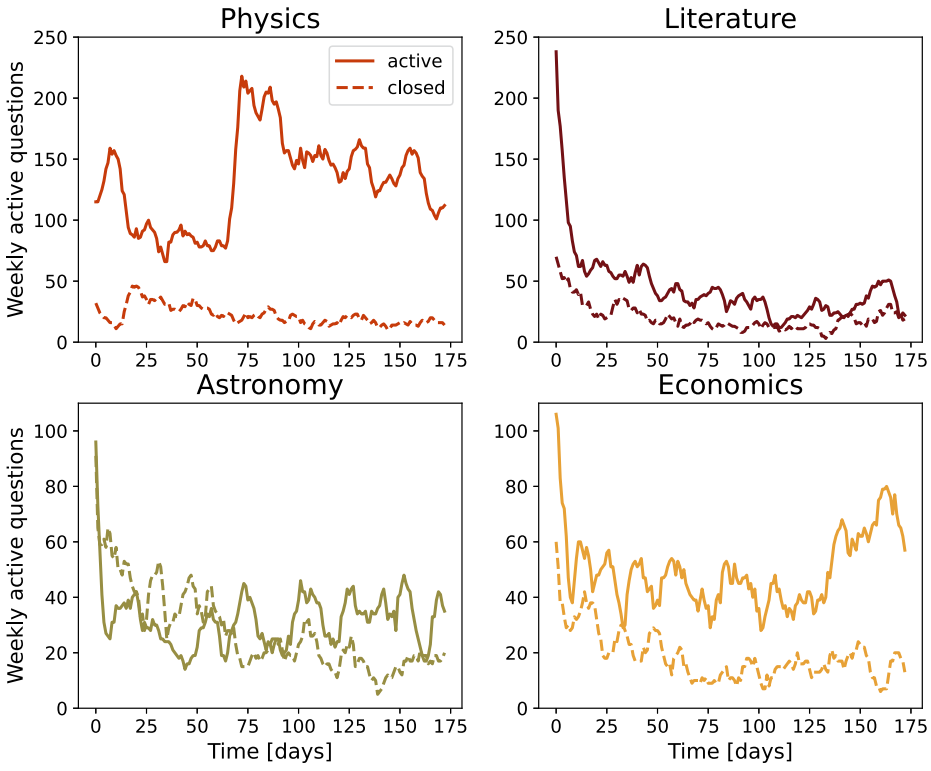

Figure 2 Variations in the number of active questions in SE communities. Number of active questions within 7 days sliding windows on the four pairs of Stack Exchange websites: Astronomy, Literature, Economics and Physics. Solid lines – active sites; dashed lines – closed sites

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:

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

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 $\{\boldsymbol{p}\}_{2x2}$ 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, $\boldsymbol{p}$ given network G is proportional to the likelihood of generating network G, $P(G|\theta, \boldsymbol{p})$, prior on SBM matrix $P(\boldsymbol{p})$ and prior on block assignments $P(\theta)$:

$$P(\theta, \boldsymbol{p}|G) = P(G|\boldsymbol{p}, \theta)P(\boldsymbol{p})P(\theta), \tag{2}$$

The likelihood of generating a network G is defined as:

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

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 $\boldsymbol{p}$ is the uniform distribution over all block matrices whose elements satisfy the constraint for the core-periphery structure $0 < p_{22} < p_{12} < p11 < 1$. 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 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:

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

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,

$$\Delta_n = \frac{t_n - t_{n-1}}{t_a}. \tag{5}$$

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:

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

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 $I_{bn} = 1$, 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

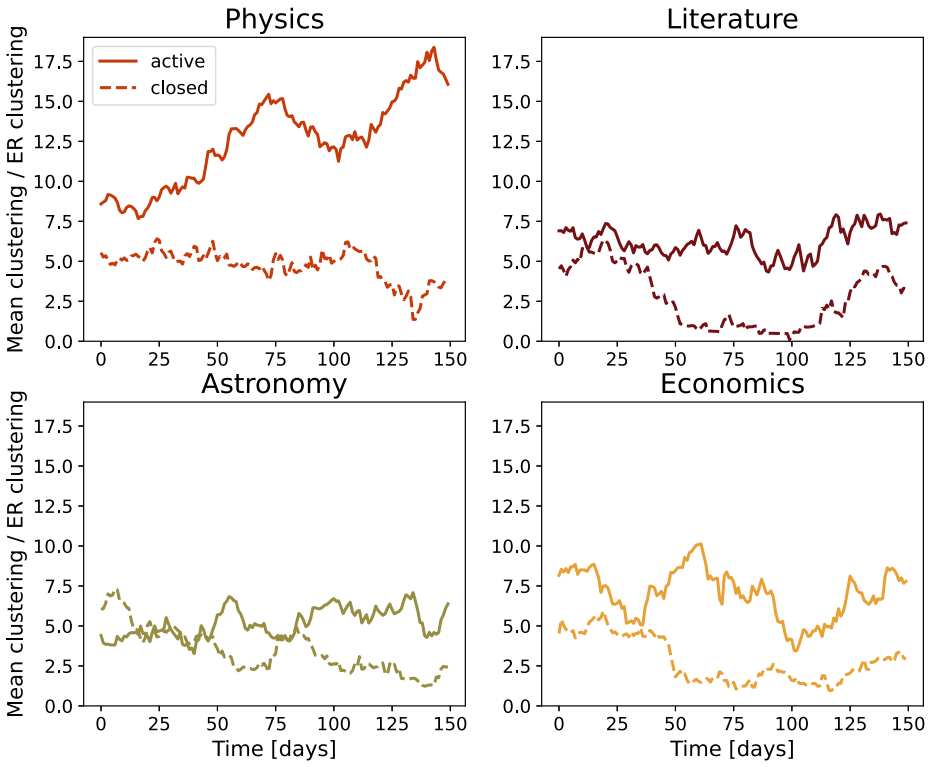

Figure 3 Normalized mean clustering coefficient of 30 days sub-networks for four pairs of Stack Exchange websites: Astronomy, Literature, Economics, and Physics. Solid lines – active sites; dashed lines – closed sites

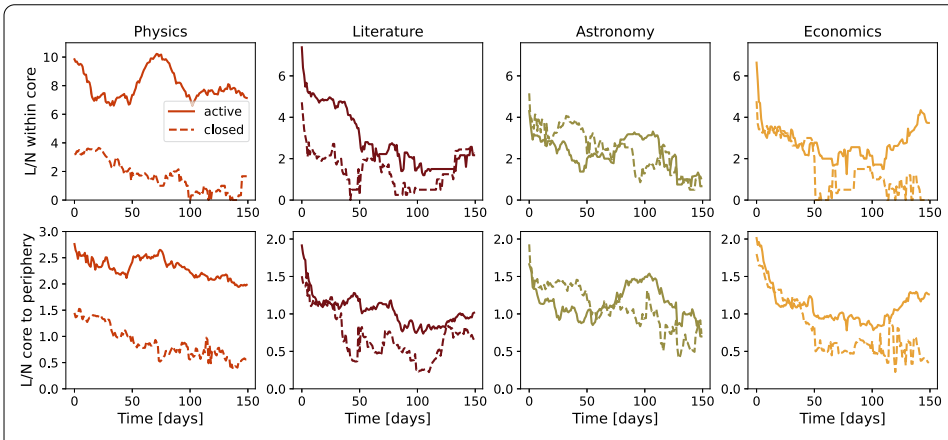

Figure 4 Connectivity among users within the core and between core and periphery. Links per node in core - top panel and links per node between core and periphery - bottom panel for the four pairs of Stack Exchange websites: Astronomy, Literature, Economics, and Physics. Solid lines – active sites; dashed lines – closed sites

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 that 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.

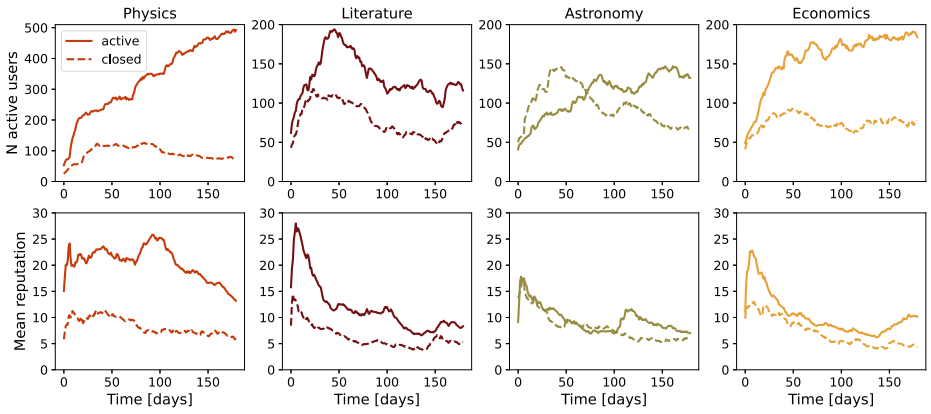

Figure 5 Active users within SE communities and their mean dynamic reputation. The number of active users (users with a reputation higher than 1) - top panel, and mean Dynamic Reputation within active users – bottom panel for the four pairs of Stack Exchange websites: Astronomy, Literature, Economics, and Physics. Solid lines – active sites; dashed lines - closed sites

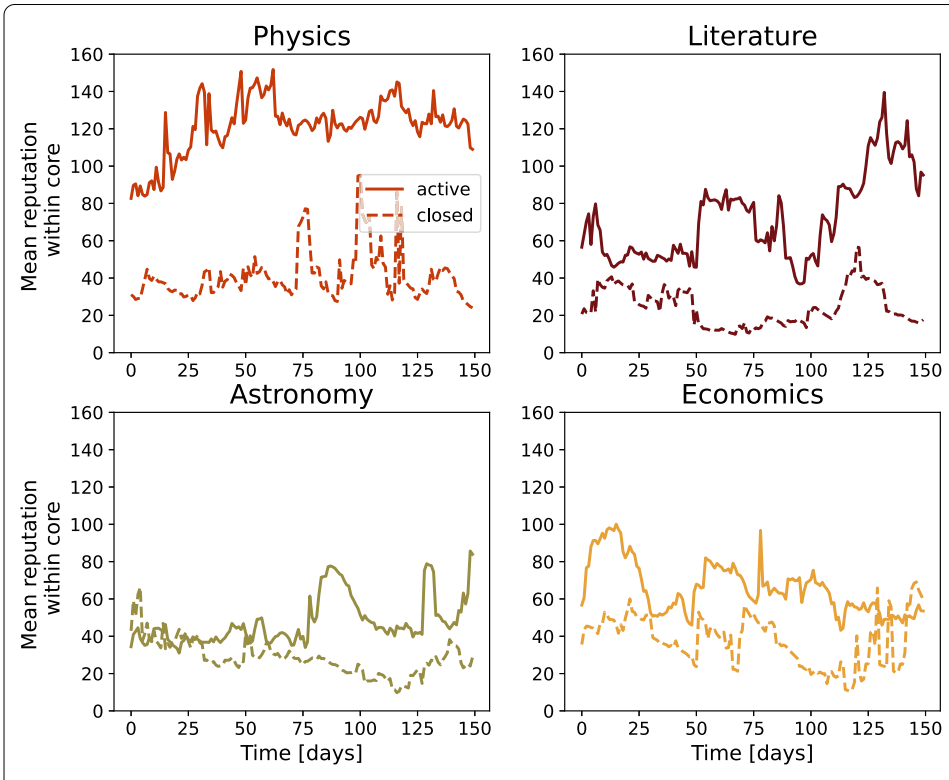

Figure 6 Mean Dynamical reputation within the core for four pairs of Stack Exchange websites: Astronomy, Literature, Economics, and Physics. Solid lines – active sites; dashed lines – closed sites

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

Publisher's Note

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

References
1. 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
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
3. 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
4. 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
5. Abrahao 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
6. 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
7. 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
8. Santos T, Walk S, Kern R, Strohmaier M, Helic D (2019) Activity archetypes in question-and-answer (q8a) websites—a study of 50 stack exchange instances. ACM Trans Soc Comput 2(1):1–23. https://doi.org/10.1145/3301612
9. 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
10. 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
11. 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
12. 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
13. Oliver PE, Marwell G (2001) Whatever happened to critical mass theory? A retrospective and assessment. Sociol Theory 19(3):292–311. https://doi.org/10.1111/0735-2751.00142
14. 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
15. 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
16. Lőrincz 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
17. 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
18. 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
19. 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, Boguná 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

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

PAPER: Interdisciplinary statistical mechanics

Universal growth of social groups: empirical analysis and modeling

Ana Vranić[1,*], Jelena Smiljanić[1,2] and Marija Mitrović Dankulov[1]

[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.

1742-5468/22/123402+19$33.00

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

Contents

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

J. Stat. Mech. (2022) 123402

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. Reddit[3] enables sharing of diverse web content, and members of this platform interact exclusively online through posts and comments. The Meetup[4] 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/.

[4] www.meetup.com.

J. Stat. Mech. (2022) 123402

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. $S_i(t) = \sum_{k=t_{i0}}^{k=t} N_i(t)$, where t_{i0} 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)

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

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 $\sigma = 3.07$.

J. Stat. Mech. (2022) 123402

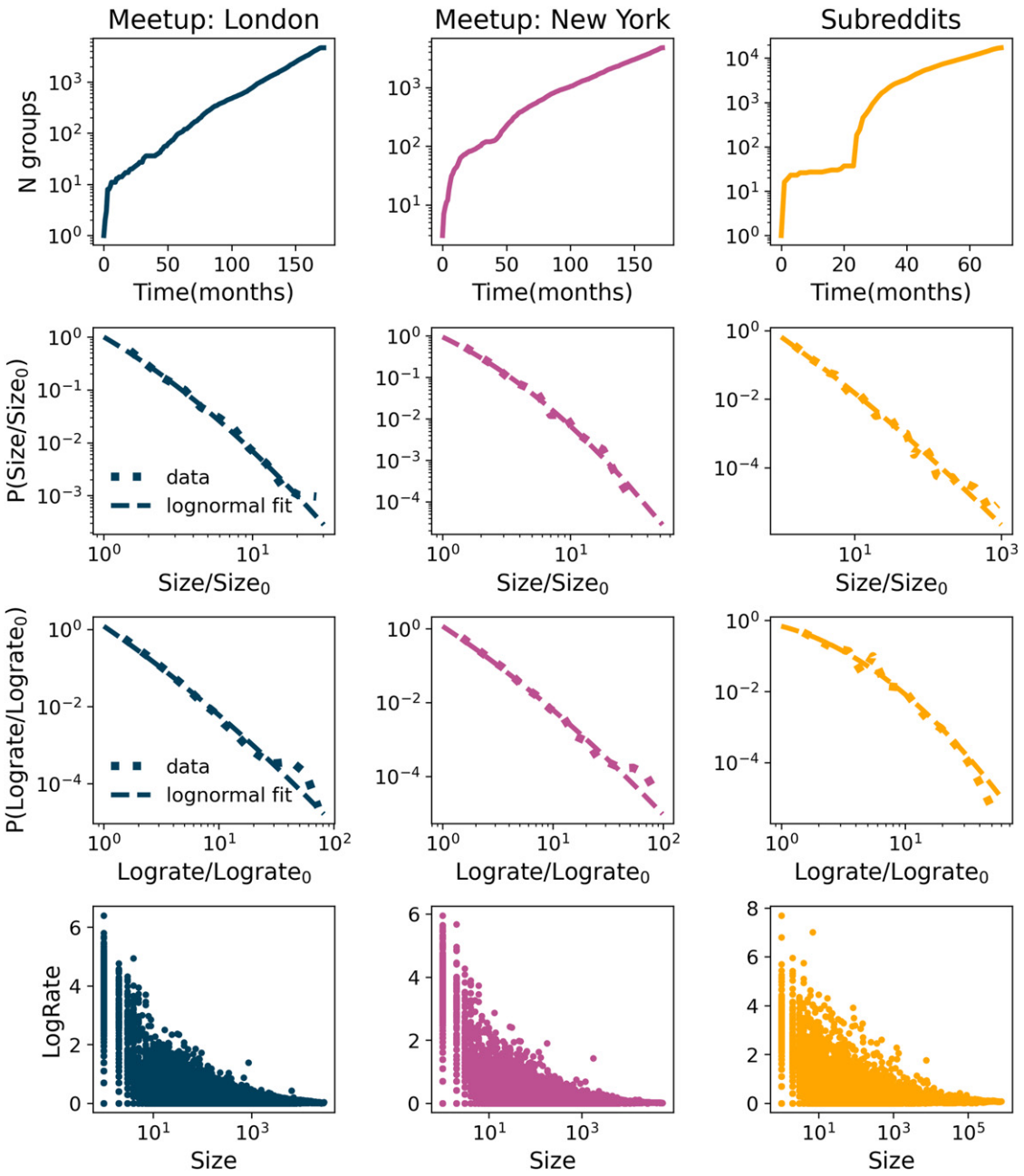

Figure 1. The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London and New York and subreddits. The number of groups grows exponentially over time, while the group size distributions,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

J. Stat. Mech. (2022) 123402

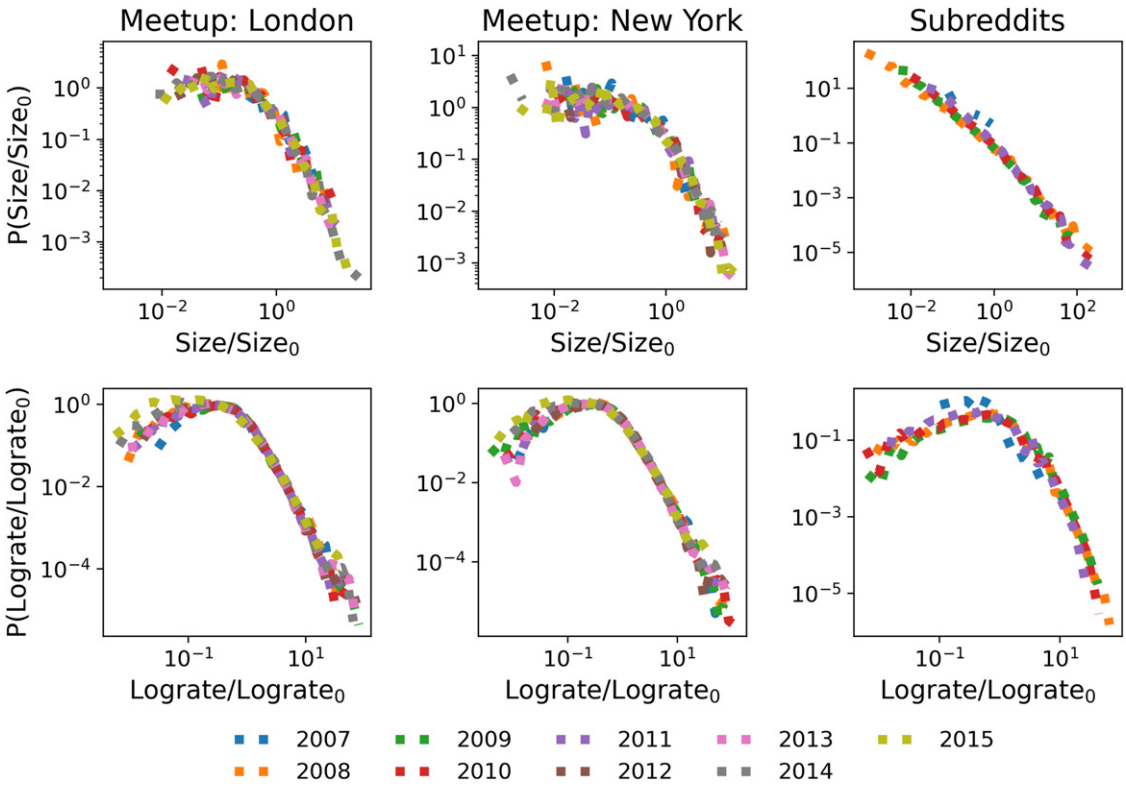

Figure 2. The figure shows the groups' sizes distributions and log-rates distributions. Figures in the top panels show the distribution of normalized sizes of groups created in the same year. Distributions for the same system and different years follow same log-normal distribution indicating existence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

J. Stat. Mech. (2022) 123402

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

J. Stat. Mech. (2022) 123402

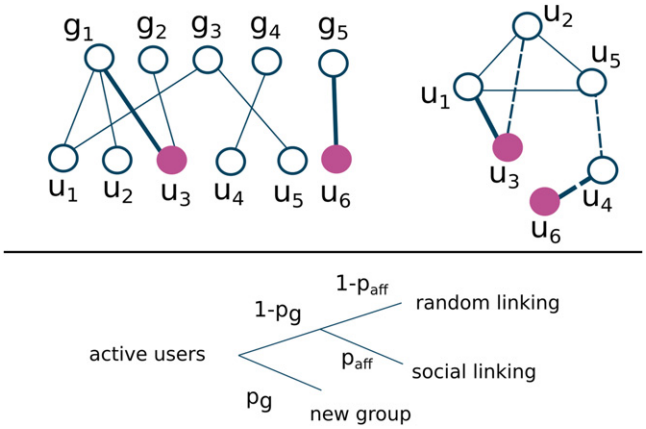

Figure 3. The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model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

J. Stat. Mech. (2022) 123402

is affiliated with group g, and zero otherwise. In matrix A, the element $A_{u_1u_2}$ equals one if members u_1 and u_2 are connected and zero otherwise. The 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

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

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.

J. Stat. Mech. (2022) 123402

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

J. Stat. Mech. (2022) 123402

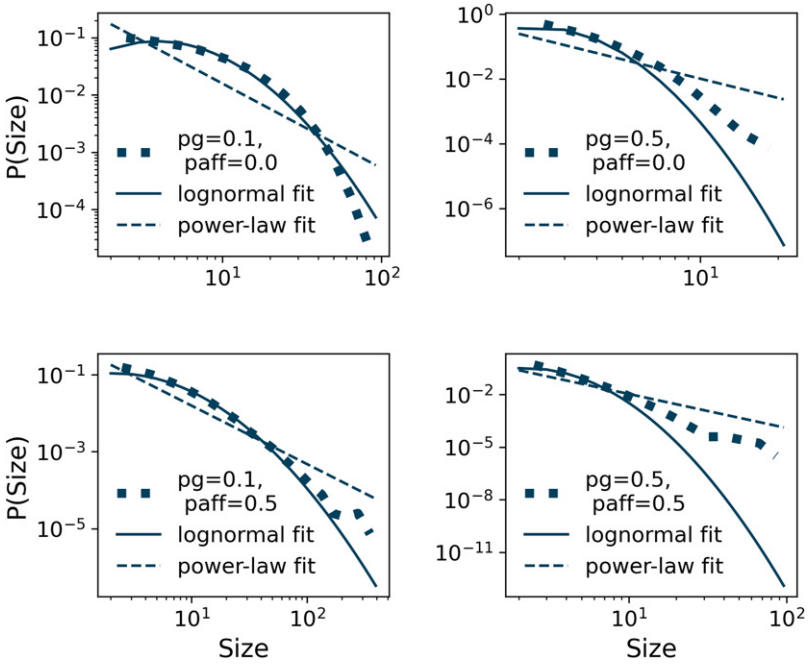

Figure 4. The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

depends on the parameter p_{aff} value. When $p_{\text{aff}} = 0$, 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

J. Stat. Mech. (2022) 123402

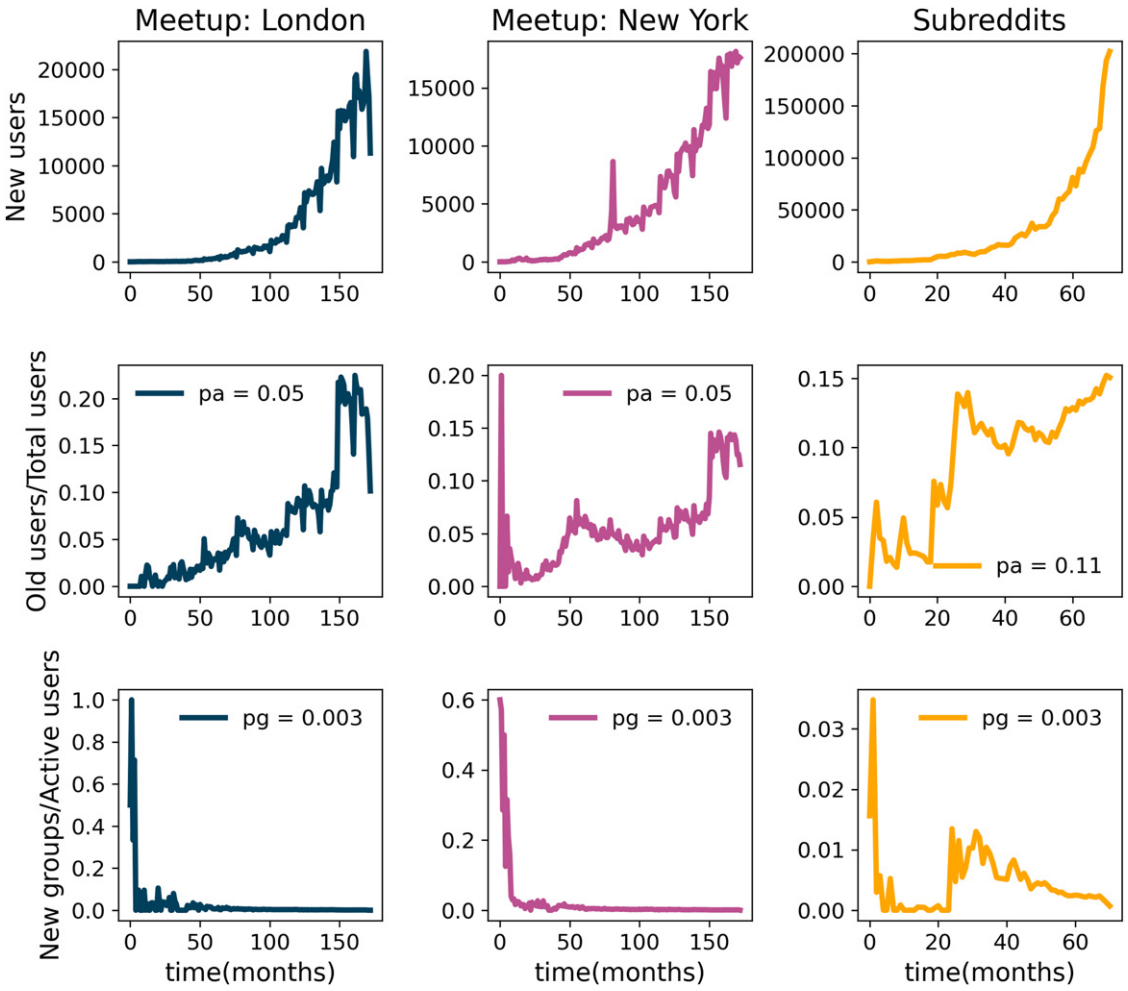

Figure 5. The time series of the number of new members (top panels). The time series of the ratio between several old active members and total members in the system (middle panels); its median value approximates the parameter p_a, the probability that the user is active. The bottom panels show the time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

J. Stat. Mech. (2022) 123402

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. $p_g(t) = \frac{Ng_{\text{new}}(t)}{N_{\text{new}}(t)+N_{\text{old}}(t)}$. 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

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

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

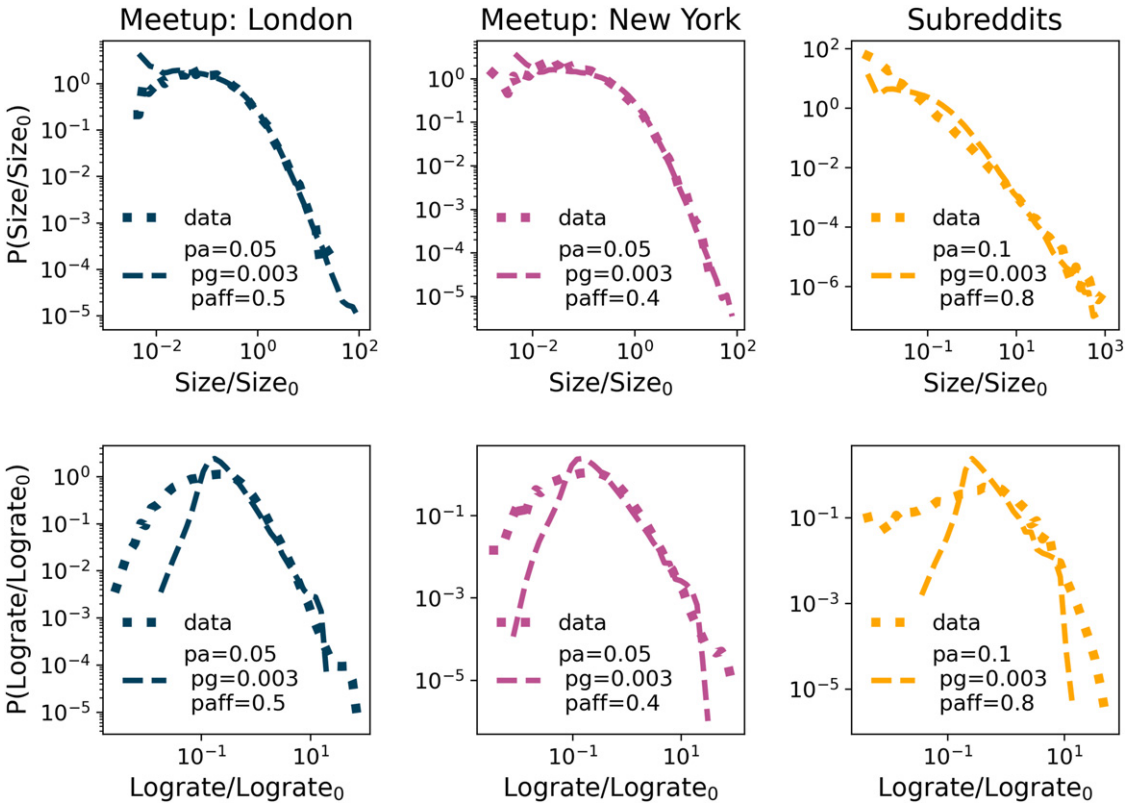

Figure 6. The comparison between empirical and simulation distribution for group sizes (top panels) and log-rates (bottom panels).

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}.

J. Stat. Mech. (2022) 123402

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.

J. Stat. Mech. (2022) 123402

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

J. Stat. Mech. (2022) 123402

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] Barthelemy 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, Haghighat 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, Iñiguez 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 evolutionProc. 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

J. Stat. Mech. (2022) 123402

[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

J. Stat. Mech. (2022) 123402

Journal of Statistical Mechanics:
Theory and Experiment

PAPER

Growth signals determine the topology of evolving networks

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

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, Pregrevia 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

Contents

*Author to whom any correspondence should be addressed.

1742-5468/20/013405+15$33.00

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

J. Stat. Mech. (2021) 013405

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].

J. Stat. Mech. (2021) 013405

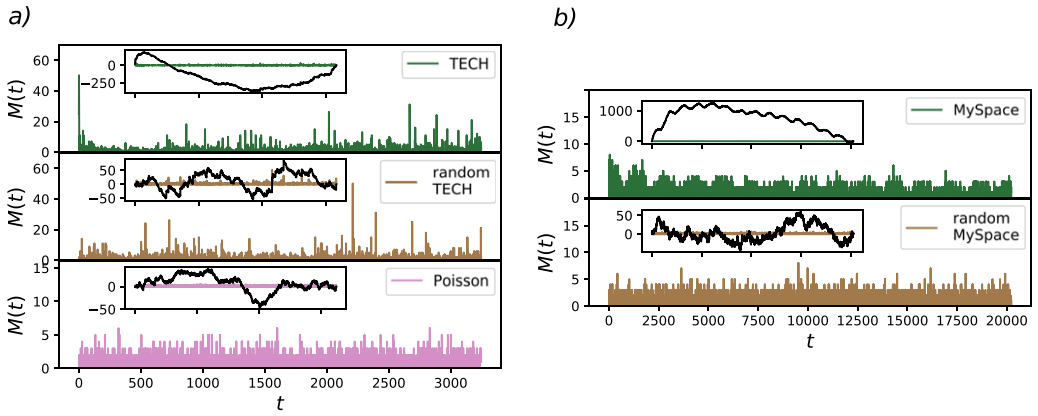

Figure 1. Growth signals for TECH (a) and MySpace (b) social groups, their randomized counterparts, and random signal drawn from Poissonian distribution with mean 1. The cumulative sums of signals' deviations from average mean value are shown in insets.

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

J. Stat. Mech. (2021) 013405

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:

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

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 $N_s = \mathrm{int}(N/s)$ non overlapping segments of length s. If N is not divisible with s the last segment will be shorter. This is handled by doing the same division from the opposite side of time series which gives us $2N_s$ segments. From each segment ν, local trend $p^m_{\nu,s}$—polynomial of order m—should be eliminated, and the variance $F^2(\nu, s)$ of detrended signal is calculated as in equation (2):

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

Then the qth order fluctuating function is:

$$\begin{aligned} F_q(s) &= \left\{ \frac{1}{2N_s} \sum_{\nu}^{2N_s} \left[F^2(\nu, s) \right]^{\frac{q}{2}} \right\}^{\frac{1}{q}}, \qquad q \neq 0 \\ F_0(s) &= \exp\left\{ \frac{1}{4N_s} \sum_{\nu}^{2N_s} \ln \left[F^2(\nu, s) \right] \right\}, \quad q = 0. \end{aligned} \tag{3}$$

The fluctuating function scales as power-law $F_q(s) \sim s^{H(q)}$ and the analysis of log–log plots $F_q(s)$ gives us an estimate of multifractal Hurst exponent $H(q)$. Multifractal signal has different scaling properties over scales while monofractal is independent of the scale, i.e., $H(q)$ is constant.

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

J. Stat. Mech. (2021) 013405

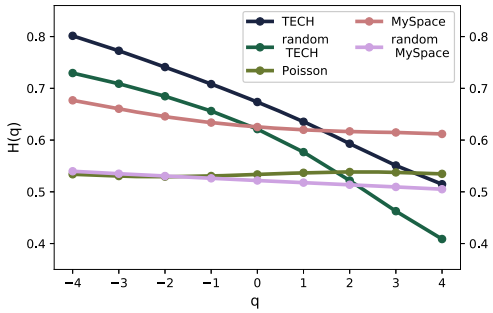

Figure 2. Dependence of Hurst exponent on parameter q for all five signals shown in figure 1 obtained with MFDFA.

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 MFDFA, 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

J. Stat. Mech. (2021) 013405

chosen according to probability

$$\Pi_i(t) \sim k_i(t)^{\beta}\tau_i^{\alpha} \tag{4}$$

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 $M \geqslant 1$ new nodes to the network and link them to $L \geqslant 1$ 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 $\beta(\alpha^*)$, $L \geqslant 2$, 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 $-\infty < \alpha \leqslant 0$ and $\beta \geqslant 1$ and constant L.

J. Stat. Mech. (2021) 013405

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 $-3 \leqslant \alpha \leqslant -0.5$ and $1 \leqslant \beta \leqslant 3$ 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]

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

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 $\{P_1, \ldots, P_N\}$ 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 $J(P_1, \ldots, P_N)$ 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 NND $= 0$ and can not be compared using NND. For these reasons, authors also introduce average node distance distribution of a graph $\mu(G) = \{\mu(1), \ldots, \mu(d)\}$, 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

J. Stat. Mech. (2021) 013405

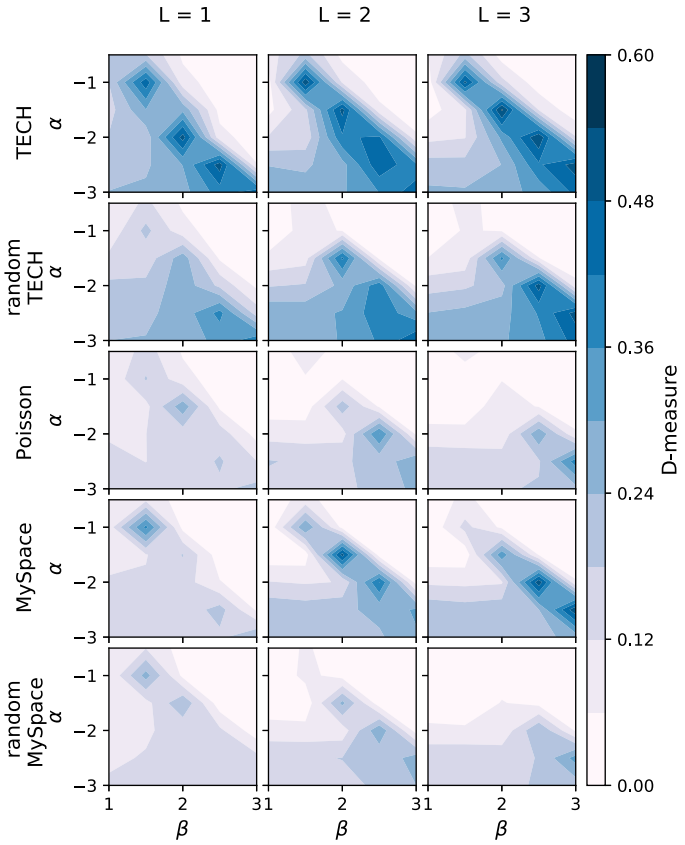

Figure 3. The comparison of networks grown with growth signals shown in figure 1 versus ones grown with constant signal $M = 1$, for value of parameter $\alpha \in [-3, -1]$ and $\beta \in [1, 3]$. $M(t)$ is the number of new nodes, and L is the number of links added to the network in each time step.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

J. Stat. Mech. (2021) 013405

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

J. Stat. Mech. (2021) 013405

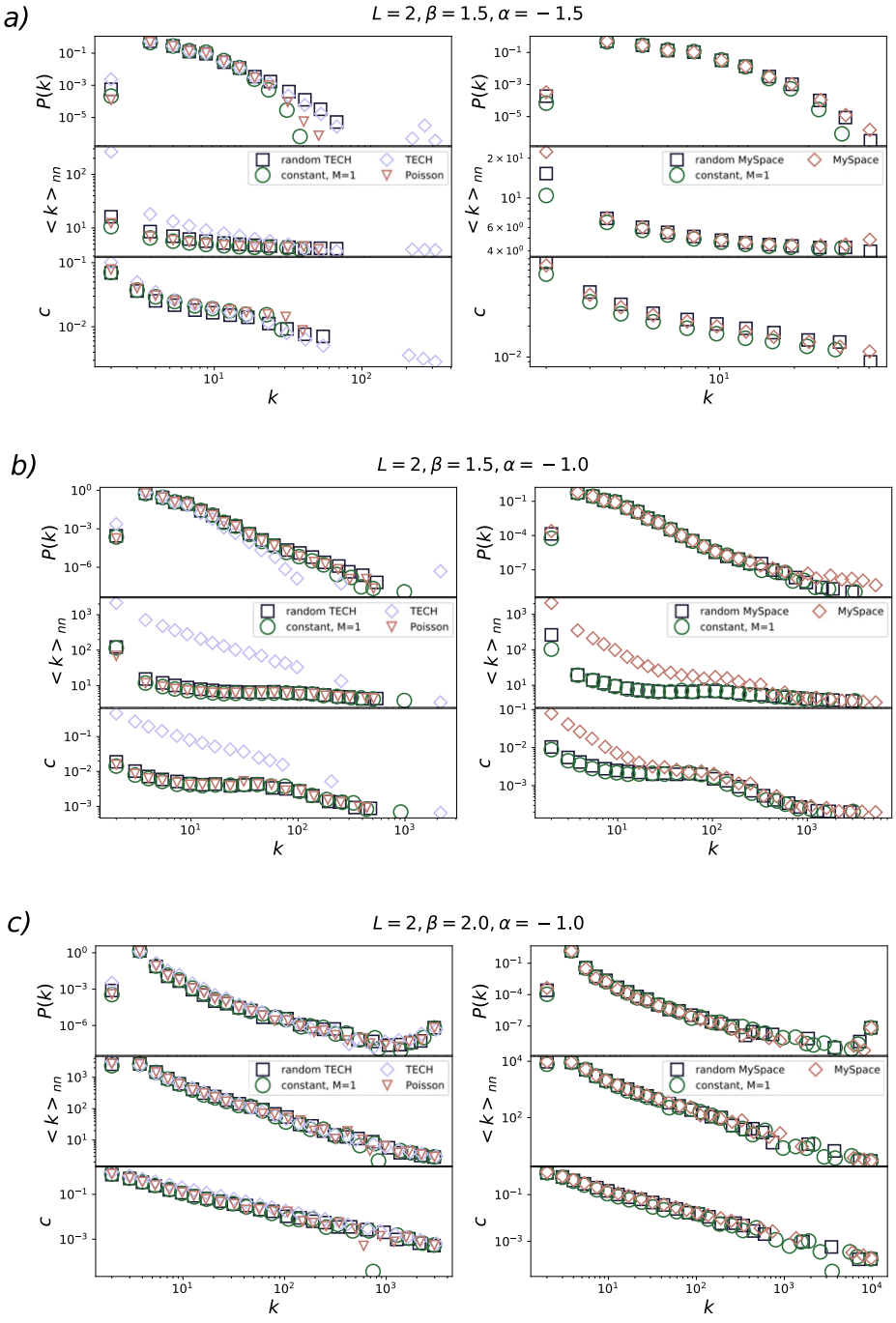

Figure 4. Degree distribution, the dependence of average first neighbor degree on node degree, dependence of node clustering on node degree for networks grown with different time-varying and constant signals. Model parameters have the values $\alpha = -1.5$, $\beta = 1.5$ (a), $\alpha = -1.0$, $\beta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

J. Stat. Mech. (2021) 013405

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 $\alpha \leqslant \alpha^*$, 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 Erdös–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

J. Stat. Mech. (2021) 013405

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 attache 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 gels 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.

J. Stat. Mech. (2021) 013405

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, Barthelemy 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, Dolenšek 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

J. Stat. Mech. (2021) 013405

[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, Mitrovi M, Gligorijevi V and Tadi 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

J. Stat. Mech. (2021) 013405

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

A. Vranić,[1] J. Vučičević,[1] J. Kokalj,[2,3] J. Skolimowski,[3,4] R. Žitko,[3,5] J. Mravlje,[3] and D. Tanasković[1]

[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

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

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

where t is the hopping between the nearest neighbors on either triangular or square lattice. $c^{\dagger}_{i\sigma}$ 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

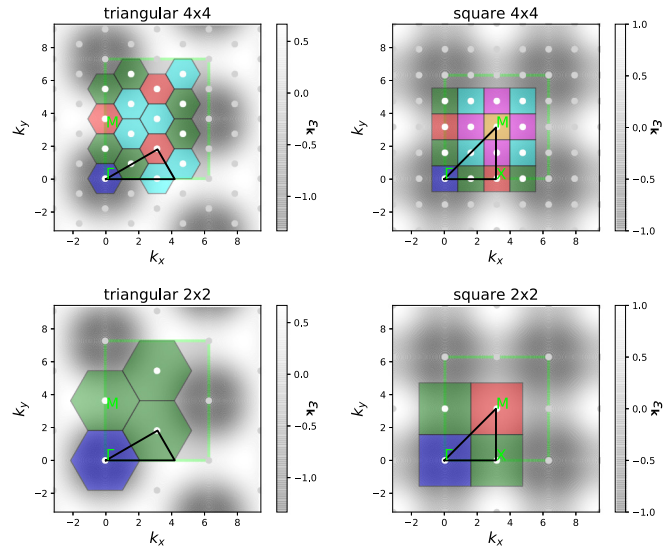

FIG. 1. DCA patches in the Brillouin zone. The irreducible Brillouin zone is marked by the black triangle. The dispersion relation is shown in gray shading. Note the position of the Γ point in the center of the first Brillouin zone which is not marked in this figure.

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

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

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

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

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 $E_{\mathrm{tot}} = E_{\mathrm{kin}} + E_{\mathrm{pot}}$. The specific heat $C = dE_{\mathrm{tot}}/dT|_n$ is obtained by interpolating $E_{\mathrm{tot}}(T)$ 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,

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

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

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

where $X(\varepsilon) = \frac{1}{N}\sum_{\mathbf{k}} \left(\frac{\partial\varepsilon_k}{\partial k_x}\right)^2 \delta(\varepsilon - \varepsilon_k)$ is the transport function, $A(\varepsilon,\nu) = -\frac{1}{\pi}\mathrm{Im}[\nu + \mu - \varepsilon - \Sigma(\nu)]^{-1}$, 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.

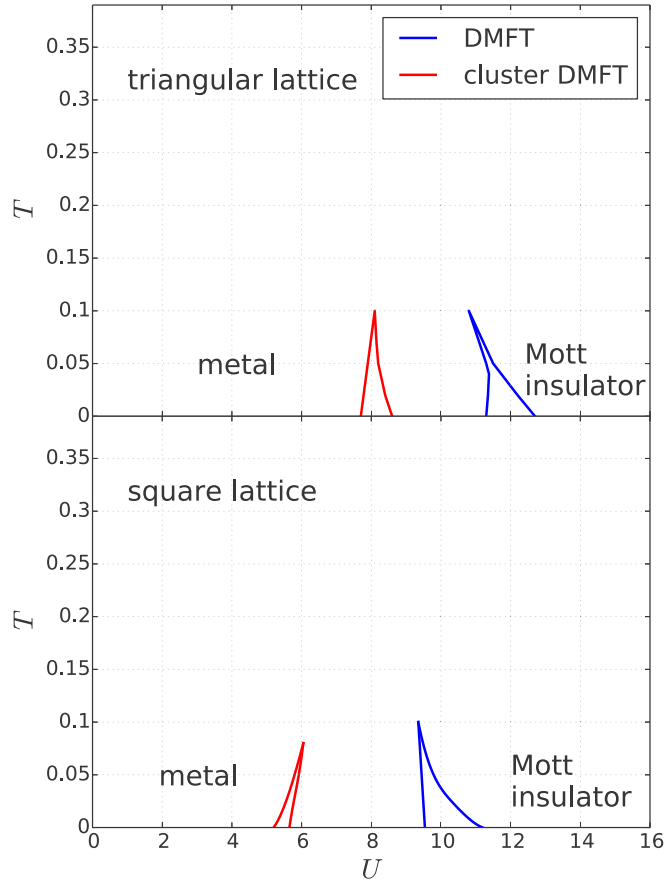

FIG. 2. Sketch of the paramagnetic phase diagram at half-filling, adapted from Refs. [46,47]. There is a region of the coexistence of metallic and insulating solution below the critical end point at T_c. The critical interaction is smaller in the cluster DMFT solution. Above T_c there is a gradual crossover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

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 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 Im Σ results presented in Fig. 3 we estimate by looking at the difference in Im Σ between the last two iterations of the cluster DMFT loop. We monitor all **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

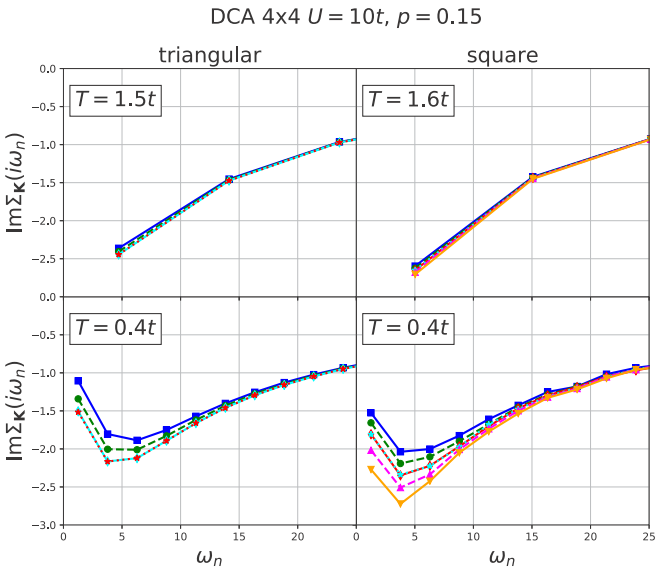

FIG. 3. Imaginary part of the self-energy at the Matsubara frequencies at different patches of the Brillouin zone for several temperatures for $p = 0.15$ hole doping. The position of the patches is indicated by the same colors as in Fig. 1. The solid lines are guide to the eye.

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 $C = dE_{\mathrm{tot}}/dT|_n$, 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 $T \geqslant 0.05t$ and cluster DMFT for $T \geqslant 0.2t$. 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

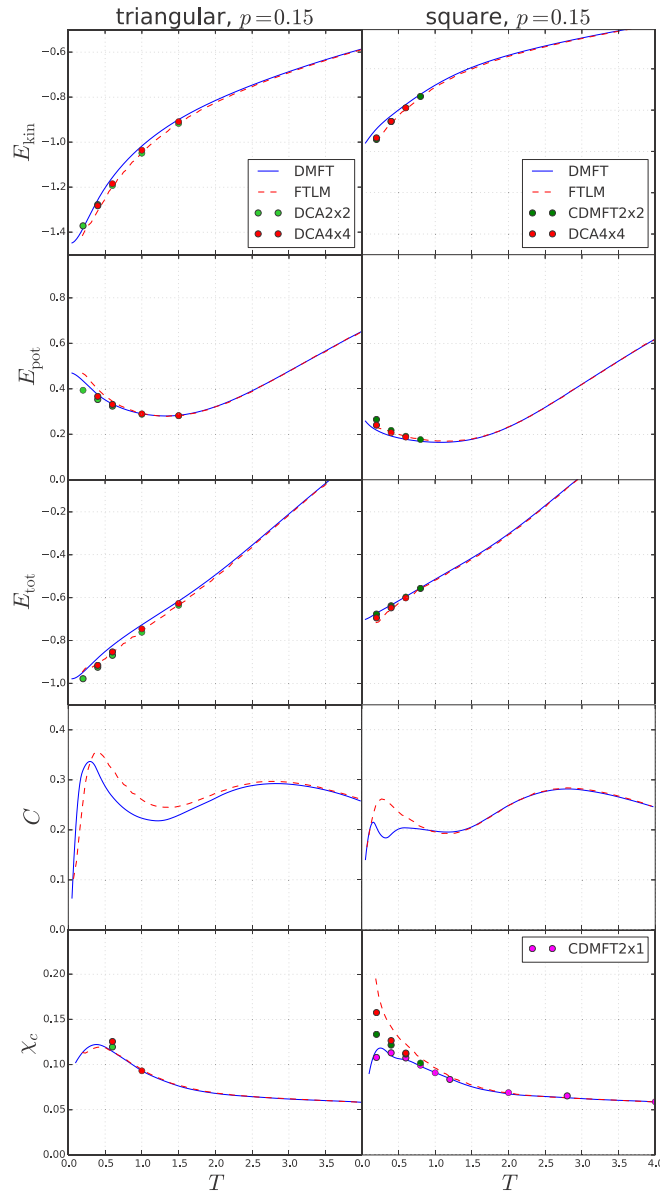

FIG. 4. Kinetic, potential, total energy, specific heat, and charge susceptibility as a function of temperature for the triangular and the square lattice at 15% doping.

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 $E_{\mathrm{tot}}(T)$.] We note that the specific heat, shown here for the fixed particle density, is slightly different than the one for the fixed chemical potential $C_\mu = dE_{\mathrm{tot}}/dT|_\mu$, 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 $E_{\rm kin}$ and $E_{\rm pot}$.

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 $T \to 0$. 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.

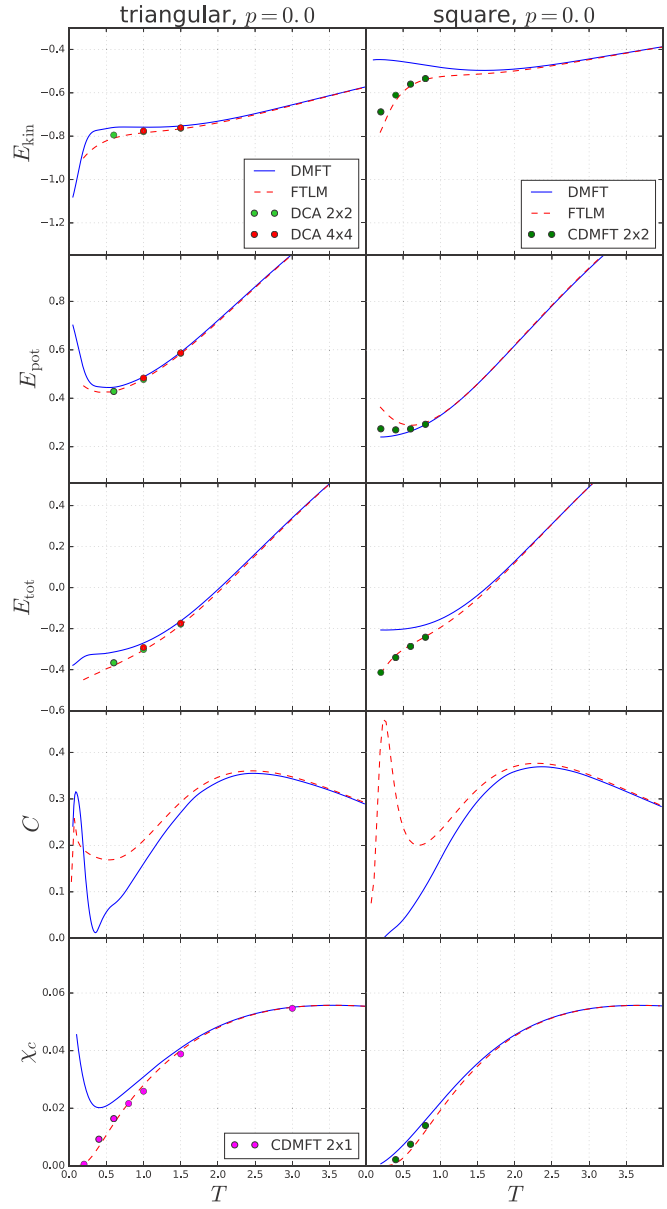

FIG. 5. Kinetic, potential, total energy, specific heat, and charge compressibility as a function of temperature for the triangular and the square lattice at half-filling.

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

$$\int_0^\infty d\omega\, \sigma(\omega) = \frac{\pi}{4V_{u.c.}}(-E_{\rm kin}), \tag{6}$$

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

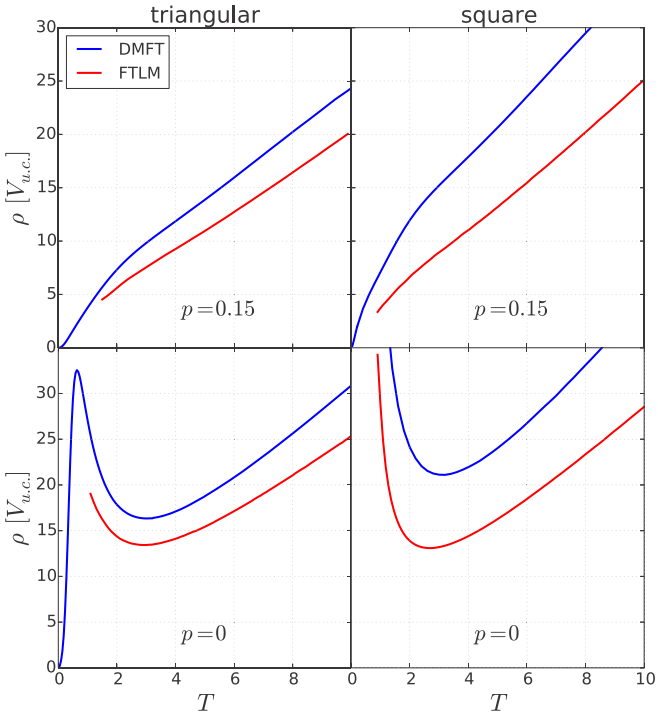

FIG. 6. Resistivity as a function of temperature.

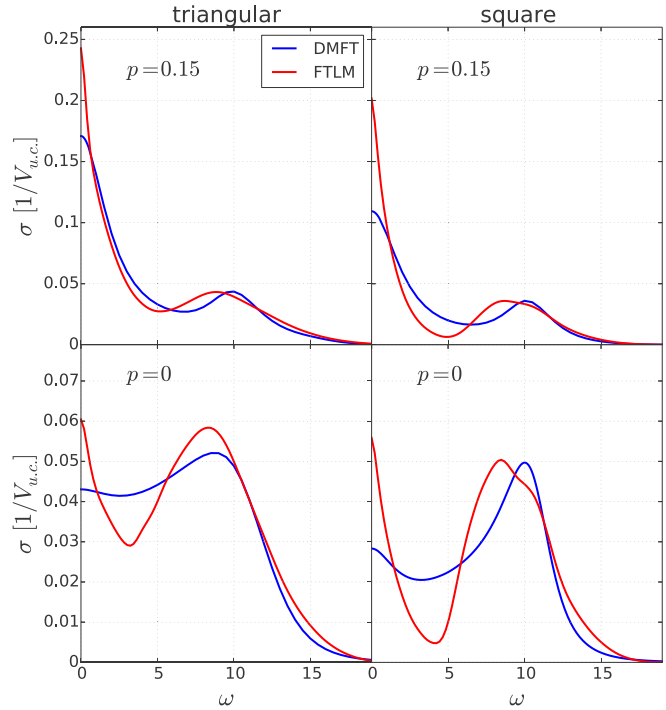

FIG. 7. Optical conductivity at $T = 1.4$.

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 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] $\rho_{\mathrm{MIR}} \sim \sqrt{2\pi} \approx 2.5$.

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]

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

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 $\omega \to 0$ (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 ∼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 $\rho = \sigma^{-1}(\omega \to 0)$.

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 $|\delta\Sigma(i\omega_0)| \approx 5\times10^{-4}$ and $|\delta G(i\omega_0)| \approx 2\times10^{-5}$ at the first Matsubara frequency at $T = t$. Such a small statistical error makes the Padé analytical continuation possible for temperatures $T \lesssim 2t$.

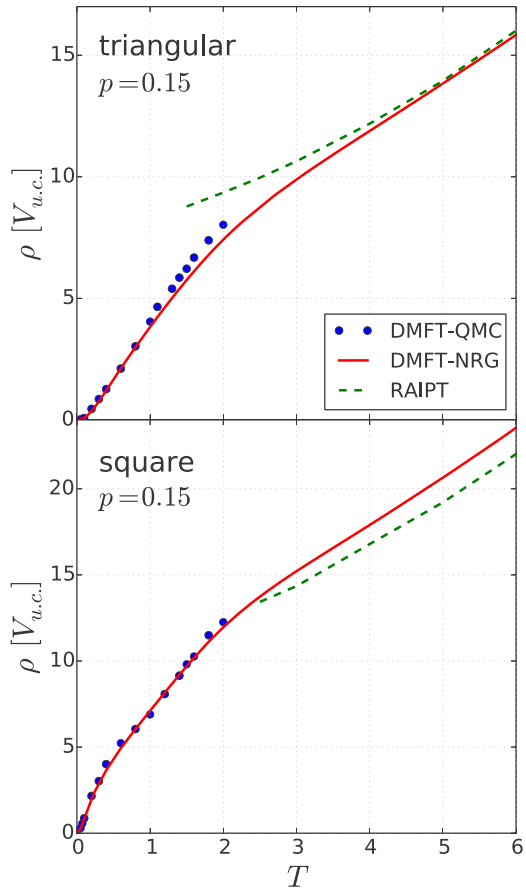

FIG. 8. DMFT-QMC (blue dots) and DMFT-NRG (red lines) resistivity as a function of temperature. The analytical continuation of the self-energy is performed with the Padé method. At high temperatures the DMFT-NRG result agrees rather well with the RAIPT (green dashed lines).

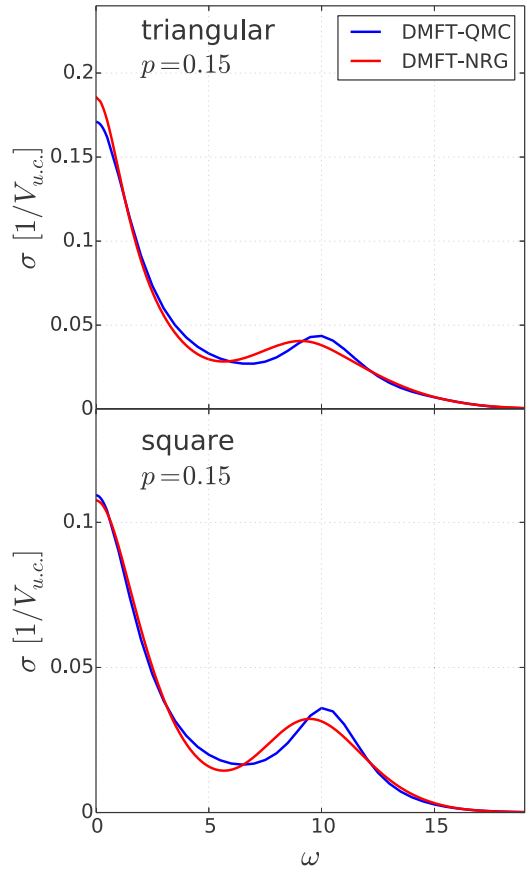

FIG. 9. DMFT-QMC and DMFT-NRG optical conductivity at $T = 1.4$.

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 than 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 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 $N_{\rm tbc} = 1,\ 4,\ 16,\ 64$, 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.

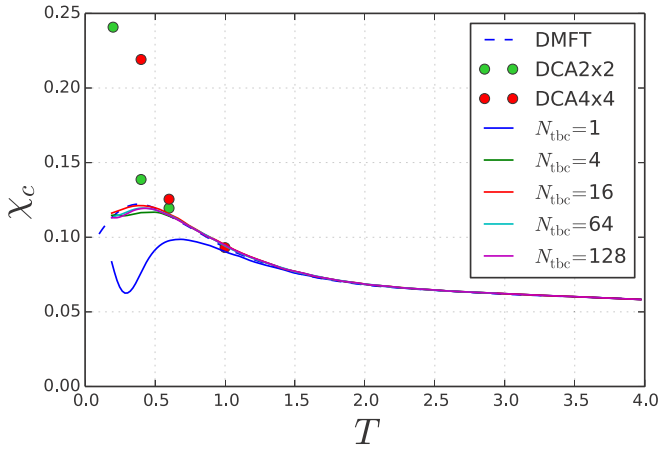

FIG. 10. Charge susceptibility as a function of temperature for the triangular lattice at $p = 0.15$ hole doping.

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.

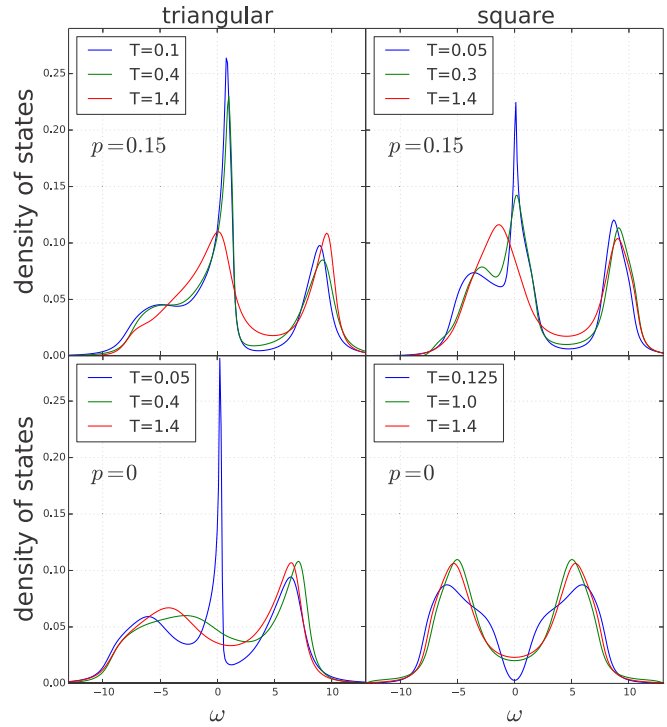

FIG. 11. Density of states in the Fermi liquid at low temperatures and in the bad-metal regime at high temperatures.

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. Vučičević, D. Tanasković, M. J. Rozenberg, and V. Dobrosavljević, Phys. Rev. Lett. **114**, 246402 (2015).

[13] H. Terletska, J. Vučičević, D. Tanasković, and V. Dobrosavljević, Phys. Rev. Lett. **107**, 026401 (2011).
[14] J. Vučičević, 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.
[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.
[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. Vučičević, 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. Prushke, 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. Vučičević 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. Ayral, 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).