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 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.

J. Stat. Mech. (2022) 123402

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 $\sigma = 3.07$.

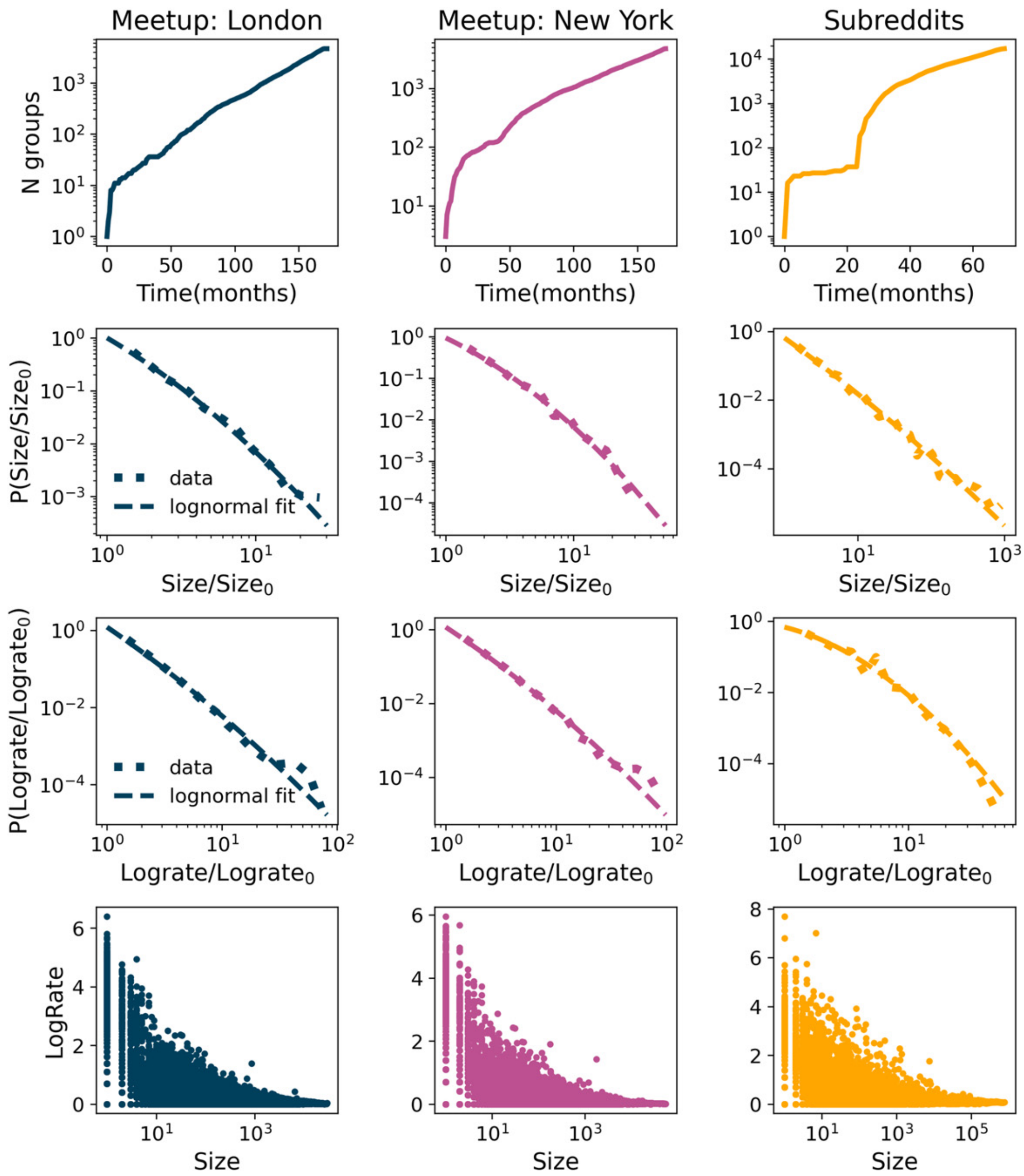

Figure 1. The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London and New York and subreddits. The number of groups grows exponentially over time, while the group size distributions, 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.

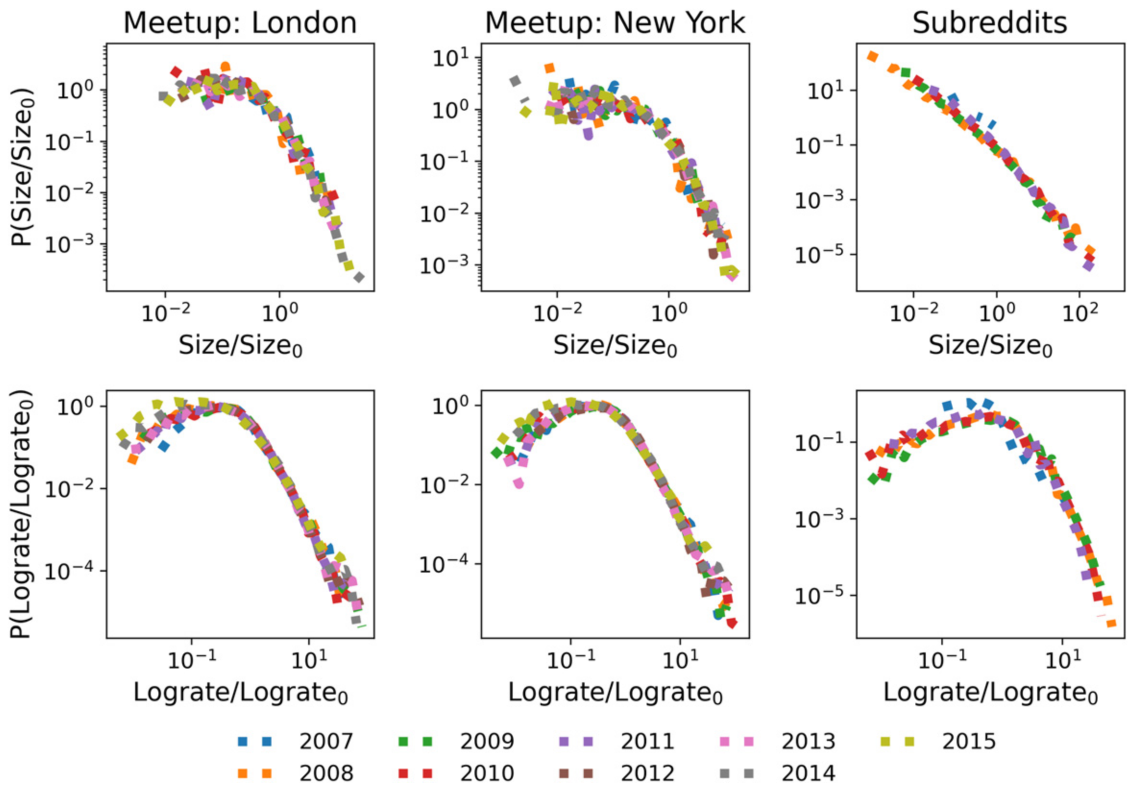

Figure 2. The figure shows the groups' sizes distributions and log-rates distributions. Figures in the top panels show the distribution of normalized sizes of groups created in the same year. Distributions for the same system and different years follow same log-normal distribution indicating existence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

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

We are considering a relatively significant period for online groups. The fast expansion of information communications technologies (ICT) changed how members access online systems. With the use of smartphones, online systems became more available, which led to the exponential growth of ICTs systems and potential change in the mechanisms that influence the social groups' growth. For these reasons, we aggregate groups according to the year they were founded for each of the three data sets and look at the distributions of their sizes at the end of 2017 for Meetup groups and 2011 for Reddit. For each year and each of the three data sets, we calculate the average size of the groups created in a year $y\langle S^y\rangle$. We normalize the size of the groups originating in year y with the corresponding average size $s_i^y = S_i^y/\langle S^y\rangle$ and calculate the distribution of the normalized sizes for each year. The distribution of normalized sizes for all years and data sets is shown in figure 2. All distributions exhibit log-normal behavior. Furthermore, the distributions for the same data set and different years follow a universal curve with the same value of parameters μ and σ. The universal behavior is observed for the distribution of normalized log-rates as well, see figure 2 (bottom panels). These results indicate that the growth of the social groups did not change due to the increased growth of members in systems. Furthermore, it implies that the growth is independent of the size of the whole data set.

4. Model

The growth of social groups cannot be explained by the simple rules of Gibrat's law. Previous research on group growth and longevity has shown that social connections with members of a group influence individual's choice to join that group [19, 25]. Individuals' interests and the need to discover new content or activity also influence the diffusion of individuals between groups. Furthermore, social systems constantly grow since new members join every minute. The properties of the growth signal that describes the arrival of new members influence both dynamics of the system [32, 33] and the structure of social interactions [34]. The number of social groups in the social systems is not constant. They are constantly created and destroyed.

In [19], the authors propose the co-evolution model of the growth of social networks. In this model, the authors assume that the social system evolves through the co-evolution of two networks: a network of social contacts between members and a network of members' affiliations with groups. This model addresses the problem of the growth of social networks that includes both linking between members and social group formation. In this model, a member of a social system selects to join a group either through random selection or according to her social contacts. In the case of random selection, there is a selection preference for larger groups. If a member chooses to select a group according to her social contacts, the group is selected randomly from the list of groups with which her friends are already affiliated.

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

J. Stat. Mech. (2022) 123402

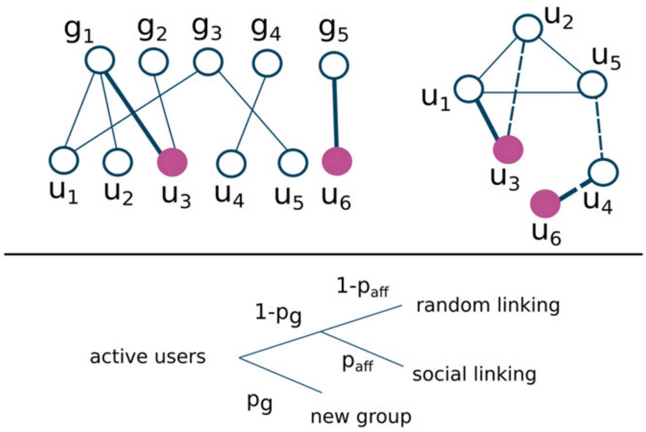

Figure 3. The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model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

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.

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

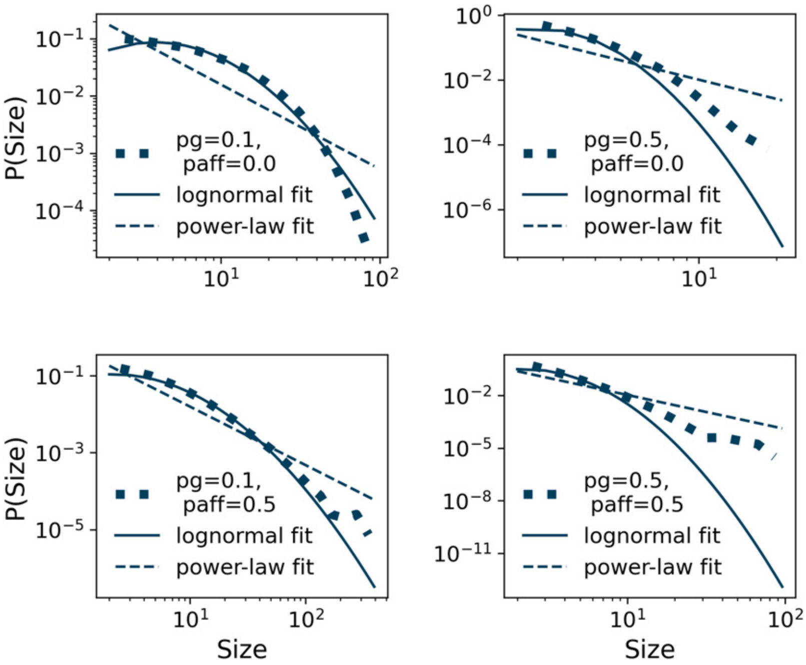

Figure 4. The distribution of sizes for different values of p_g and p_{aff} and constant p_a and growth of the system. The combination of the values of parameters of p_g and p_{aff} determine the shape and the width of the distribution of group sizes.

depends on the parameter p_{aff} value. When $p_{\mathrm{aff}} = 0$, social connections are irrelevant to the group's choice, and members select groups randomly. The obtained distribution slightly deviates from log-normal, especially for large group sizes. In this case, large group sizes become more probable than in the case of the log-normal distribution. The non-zero value of parameter p_{aff} means that the choice of a group becomes dependent on social connections. When a member chooses a group according to her social connections, larger groups have a higher probability of being affiliated with the social connections of active members, and thus this choice resembles preferential attachment. For these reasons, the obtained size distribution has more broad tail than log-normal distribution and begins to resemble power-law distribution.

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

5.2. Modeling real systems

The social systems do not grow at a constant rate. In [34], the authors have shown that features of growth signal influence the structure of social networks. For these reasons, we use the real growth signal from Meetup groups located in London and New York

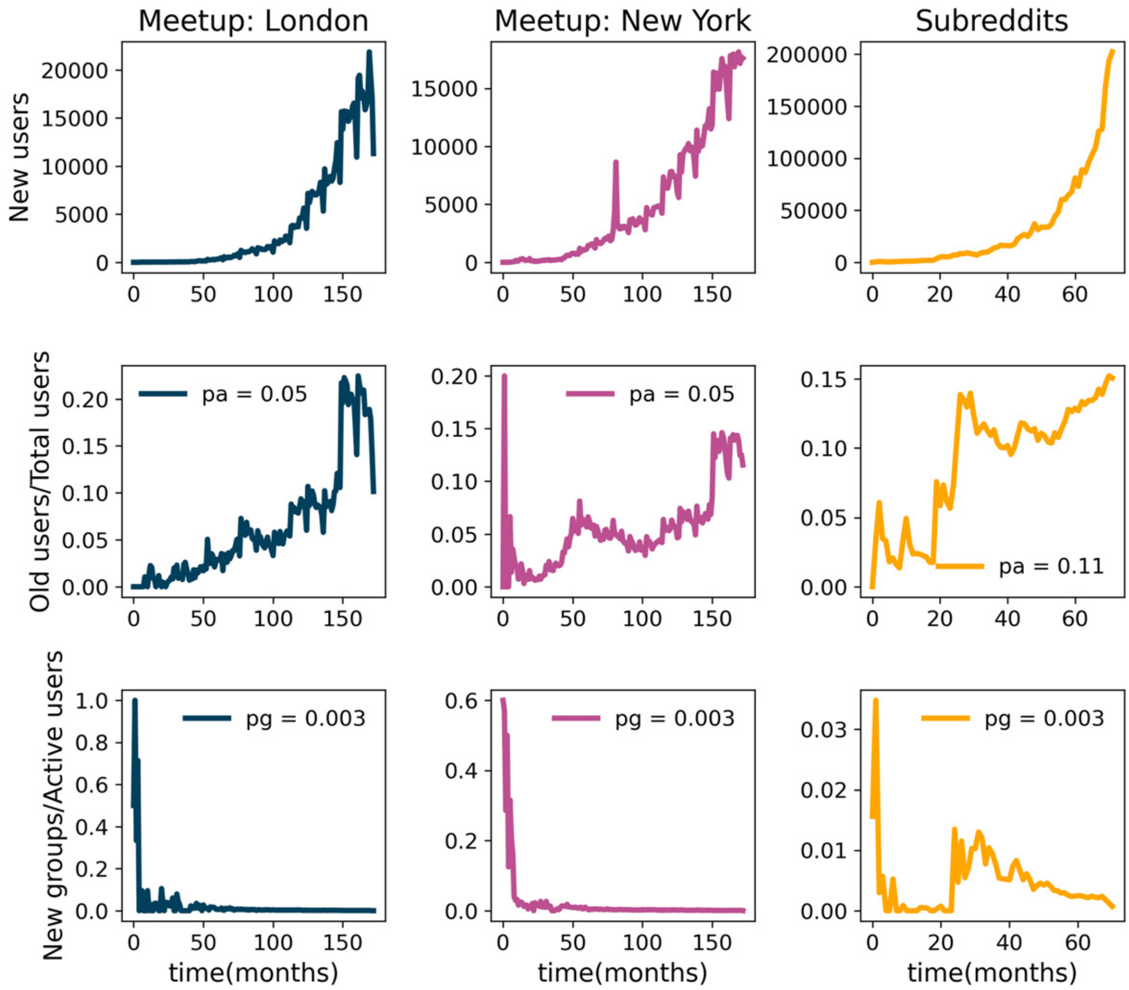

Figure 5. The time series of the number of new members (top panels). The time series of the ratio between several old active members and total members in the system (middle panels); its median value approximates the parameter p_a, the probability that the user is active. The bottom panels show the time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

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

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

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

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

groups, while its value is between 0 and 0.15 for Reddit. To simplify our simulations, we assume that p_a is constant in time and estimate its value as its median value during the 170 months for Meetup and 80 months for Reddit systems. For Meetup groups based in London and New York $p_a = 0.05$, while Reddit members are more active on average and $p_a = 0.11$ for this system.

Figure 5 bottom row shows the evolution of parameter p_g for the considered systems. The p_g in month t is estimated as the ratio between the groups created in month $tNg_{\text{new}}(t)$ and the total number of groups in that month $Ng_{\text{new}}(t) + Ng_{\text{old}}(t)$, i.e. $p_g(t) = \frac{Ng_{\text{new}}(t)}{N_{\text{new}}(t)+N_{\text{old}}(t)}$. 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

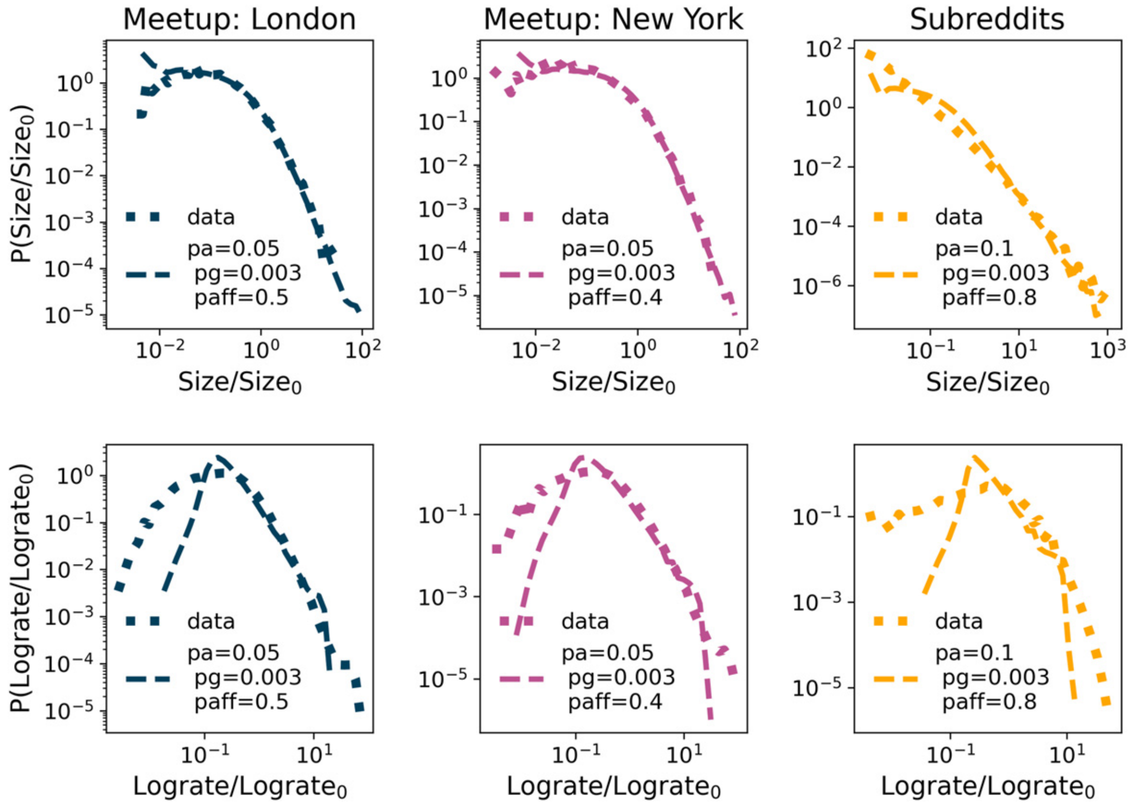

Figure 6. The comparison between empirical and simulation distribution for group sizes (top panels) and log-rates (bottom panels).

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 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

[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