Chapter 1

The growth of social groups

1.1 Social groups

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

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

The Reddit data were downloaded from https://pushshift.io/ site. This site collects posts and comments daily; data are publicly available in JSON files for each month. The selected subreddits were created between 2006 and 2011, we also filtered those active in 2017. We removed subreddits active for less than two months. The obtained dataset has 17073 subreddits with 2195677 active members. For each post, we extracted the subreddit-id, user-id and the date when the user created the post. Finally, we selected the date when each user posted on each subreddit for the first time.

1.1.1 The empirical analysis of social groups

For each Meetup group we have information when user attended the group event, while for subreddit we have detailed data about user activity, so we can extract the information when user for the first time created a post. Those dates are considered as timestamp when user joined to group. So both datasets have the same structure: (g, u, t), where t is timestamp when user u joined group g. For each

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

time step, we can calculate the number of new members in each group $N_i(t)$, and the group size $S_i(t)$. The group size at time step t is $S_i(t) = \sum_{k=t_0}^{k=t} N_i(t)$, where t_0 is month when group is created. The group size is increasing in time, as we do not have information if the user stopped to be active. Also we calculate the growth rate, as the logarithm of successive sizes $R = log(S_i(t)/S_i(t-1))$.

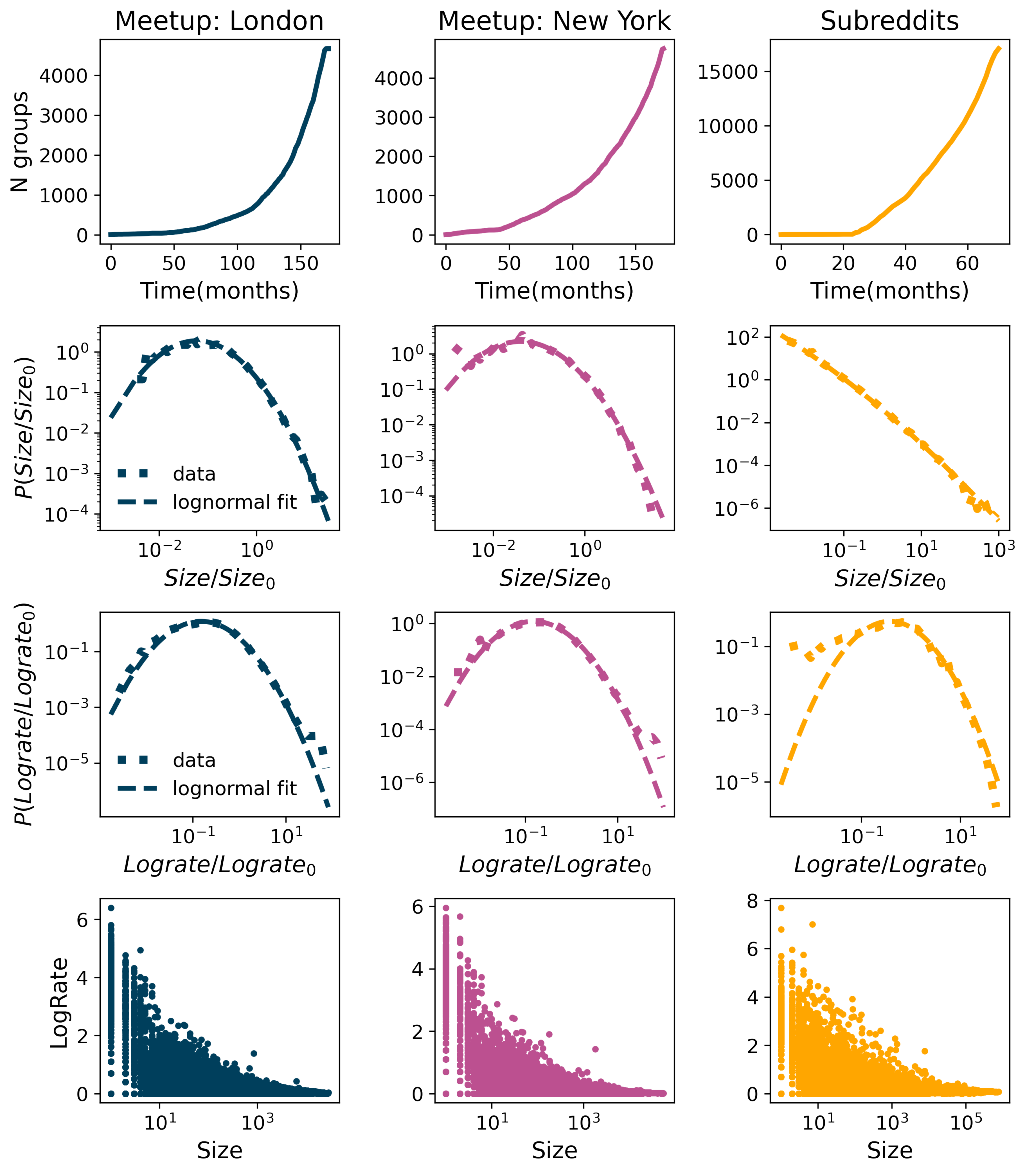

Figure 1.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011 that were active in 2017.

Even though Meetup and Reddit are different online platforms, we find some common properties of these systems; see figure 1.1. The number of groups and the number of new users grow exponentially. Still, subreddits are larger groups than Meetups. The distribution of groups sizes follows the log-normal distribution:

$$P(S) = \frac{1}{S\sigma\sqrt{2\pi}} exp(-\frac{(\ln(S)-\mu)^2}{2\sigma^2}) \tag{1.1}$$

where S is the group size and μ, and σ are parameters of the distribution. The group sizes distribution of Subreddits is a broad log-normal distribution that resembles the power law. Still, we used the loglikelihood ratio method and showed that log-normal distribution is better than the power-law. More details are given in the result section.

The simplest model that generates the lognormal distribution is multiplicative process [1]. Gibrat used this model to explain the growth of firms. The main assumption of this model is that growth rates $R = log\frac{S_t}{S_{t-\Delta t}}$ do not depend on the size S and that they are uncorrelated. Further, this imply the lognormal distribution of the sizes, while the distribution of growth rates appears to be normal distribution, [2], [3]. Figure 1.2 shows distribution of the logrates, that follow lognormal distribution, contrary to the Gibrat law. Furthermore, logrates depend on the group size 1.2. For these reasons the Gibrat law can not explain the growth of online social groups [4, 5].

The growth of online social groups has universal behavior. It is independent on the size of the group. If we aggregate the groups created in the same year y, and each group size normalize with average size $< S^y >$, $s_i^y = S_i^y / < S^y >$ we will find that group sizes distributions for the same dataset and different years fall on the same line, figure 1.2. The same characteristics are observed for the distribution of the normalized logrates 1.2. The growth is universal in time, and the group sizes distribution do not change from year to year.

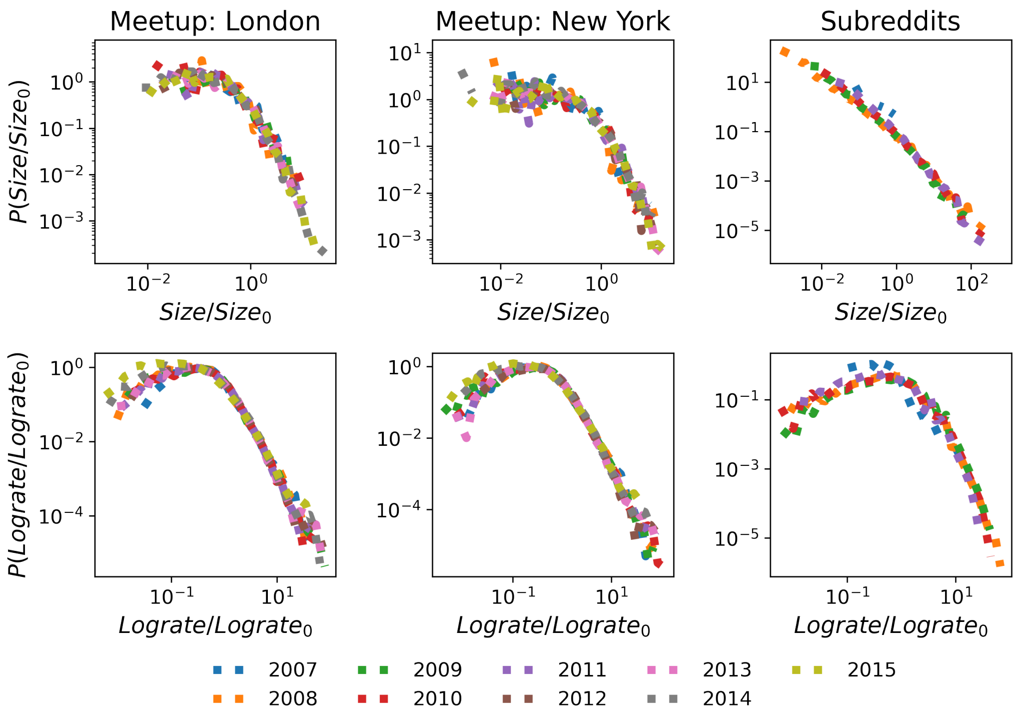

Figure 1.2: The figure shows the groups' sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

1.2 The model

The Meetup and Reddit engage members for different activities. Still, there are same underling processes for both systems. Each member can create new groups and join existing ones. Both systems grow in the number of groups and the number of users, and each user can belong to arbitrary number of groups. In the previous section we identified the universal patterns in the growth of social groups, but it appears that the growth can not be modeled with Gibrat law. The complex network models allow us to simulate the growth of these systems considering all types of the member's activities. Varying different linking rules we can identify how model parameters shape the growth process. When it comes to the user's choice of the group, it was shown that social connections play important role [6, 7]. On the other hand, user can be driven with personal interest. Diffusion between groups could be also enhanced with rich-get-richer phenomena, where users tend to join larger groups. With complex network model we can easily incorporate the nonlinear growth in the number of users and groups, as it is important parameter that shape the structure and dynamics of the complex network [8, 9, 10].

The evolution of the social groups has been studied using co-evolution model in the reference [7]. This model consists of two evolving networks: the bipartite network, that stores connections between users and groups and affiliation network of social connections. At each time step active users create new connections in the affiliation network; i.e. they make new friends. They also join existing groups or create new one, which updates the bipartite network. The group selection can be random with probability proportional to the group size, otherwise group is selected through social contacts. Using this model, authors have been able to reproduce the power-law group size distribution found in several communities, such as Flickr or LiveJournal. The empirical analysis of Meetup and Reddit groups showed that group size distribution can be log-normal, meaning that there are some different mechanisms that control the growth of the groups.

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

The schema of the model is shown on figure 1.3

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.

Figure 1.3 shows a schematic representation of our model. Similar to co-evolution model [7], we represent social system with two evolving networks, see Fig. 1.3. One network is bipartite network which describes the affiliation of individuals to social groups $\mathcal{B}(V_U, V_G, E_{UG})$. This network consists of two partitions, members V_U and groups V_G, and set of links E_{UG}, where a link $e(u, g)$ between a member u and a group g represents the member's affiliation with that group. Bipartite network grows through three activities: arrival of new members, creation of new groups, and through members joining groups. By definition, in bipartite networks links only exist between nodes belonging to different partitions. However, as we explained above, social connections affect whether a member will join a certain group or not. In the simplest case, we could assume that all members belonging to a group are connected with each other. However, previous research on this subject [11, 12, 7] has shown that the existing social connections of members in a social group are only a subset of all possible connections. For these reasons, we introduce another network $\mathcal{G}(V_U, E_{UU})$ that describes social connections between members. The social network grows through addition of new members to the set V_U and creation of new links between them. The member partition in bipartite network $\mathcal{B}(V_U, V_G, E_{UG})$ and set of nodes in members' network $\mathcal{G}(V_U, E_{UU})$ are identical.

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

In our model, the time is discrete and networks evolve through several simple rules. In each time step we add $N_U(t)$ new members and increase the size of the set V_U. For each newly added member we create the link to a randomly chosen old member in the social network G. This condition allows each member to perform social diffusion [6], i.e., to choose a group according to her social contacts. Not all members from set V_U are active in each time step. Only a subset of existing members is active in one time step. Activity of old members is a stochastic process and is determined by parameter p_a; every old member is activated with probability p_a. Old members activated in this way and new members make a set of active members $\mathcal{A}_U$ at time t.

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

The group partition V_G grows through creation of new groups. A group is created by an active user. Not all users from set V_U are active in each time step. Only a subset of existing users is active in one time step. Activity of old users is a stochastic process and is determined by parameter p_a; every old user is activated with probability p_a. Old users activated in this way and new users make a set of active users $\mathcal{A}_U$ at time t. Each active user $u \in \mathcal{A}_U$ can decide with probability p_g to create a new group, or to join an already existing group with probability $1 - p_g$.

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

$$s_{ug} = \sum_{u_1 \in \mathcal{N}_g} A_{uu_1} \tag{1.2}$$

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

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

The evolution of bipartite and social networks, and consequently growth of social groups, is

determined by parameters p_a, p_g and p_{aff}. Parameter p_a determines the activity level of members and takes values between 0 and 1. Higher values of p_a result in higher number of active members and thus faster growth of number of links in both networks, as well as the size and number of groups. Parameter p_g in combination with parameter p_a determines the growth of the set V_G. $p_g = 1$ means that members only create new groups, and the existing network consists of star-like subgraphs with members being a central nodes and groups as leafs. On the other hand $p_g = 0$ means that there is no creation of new groups and the bipartite network only grows through addition of new members and creation of new links between members and groups.
Parameter p_{aff} is especially important. It determines the importance of social diffusion. $p_{aff} = 0$ means that social connections are irrelevant and the choice of group is random. On the other hand, $p_{aff} = 1$ means that only social contacts become important for group selection.

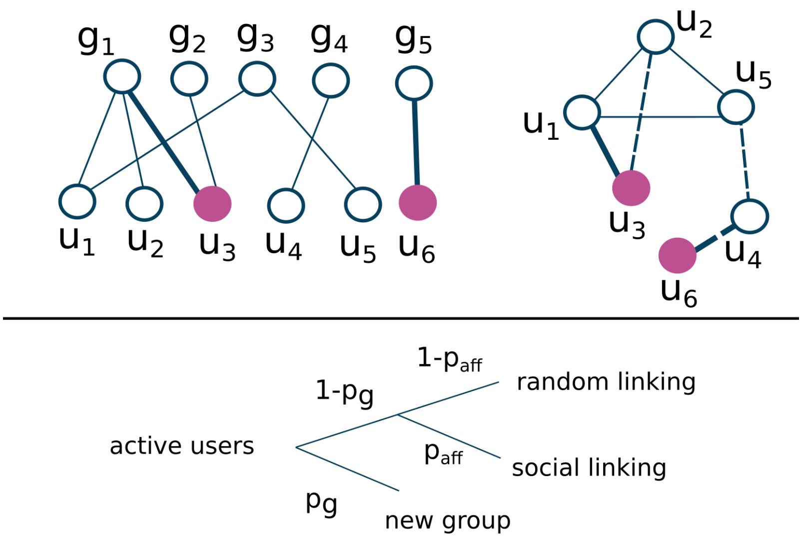

Figure 1.3: The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model schema, where p_g is probability that user create new group, while p_{aff} is probability that group choice depends on the social connections.

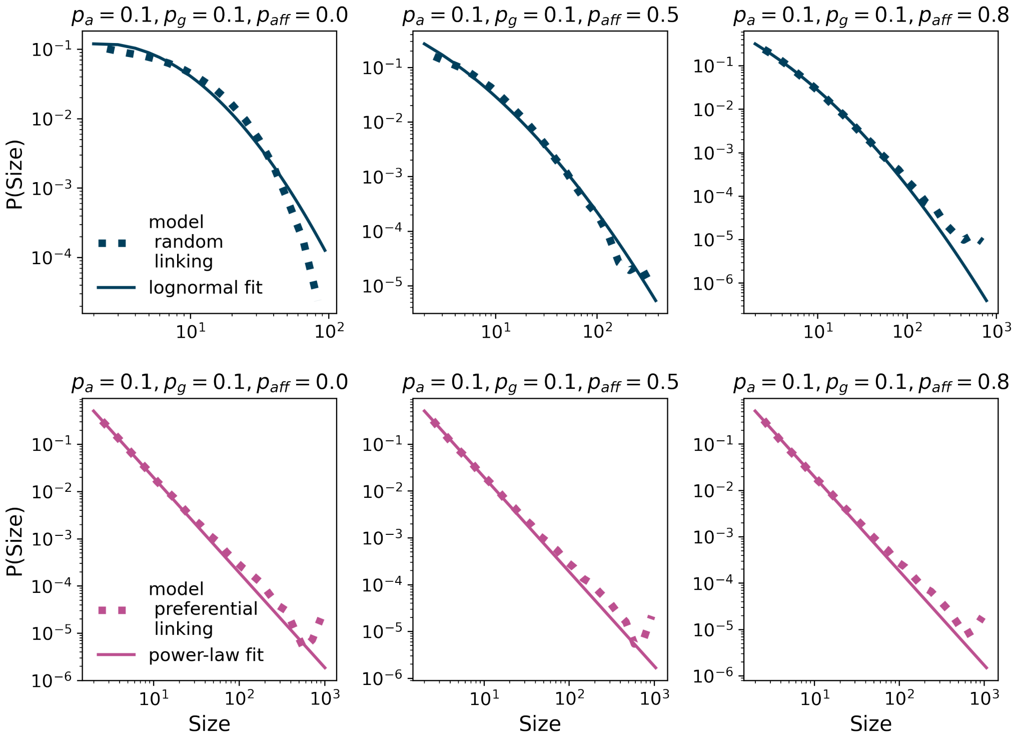

Figure 1.4: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff}. With probability, $1 - p_{aff}$, users choose a group randomly. The group sizes distribution (top row) is described with a log-normal distribution. With higher affiliation parameter, p_{aff}, distribution has larger width. The bottom row presents the case where with probability $1 - p_{aff}$ users have a preference toward larger groups. For all values of parameter p_{aff}, we find the power-law group sizes distribution.

1.3 Results

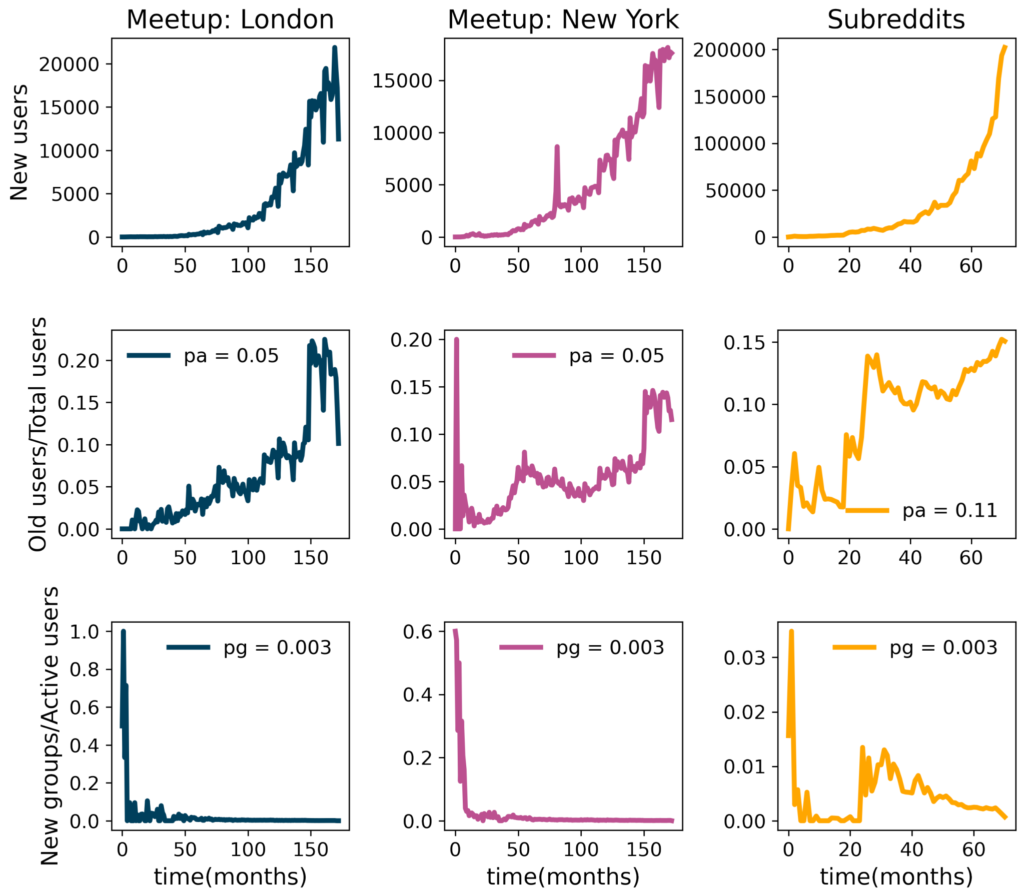

Figure 1.5: The time series of number of new members (top panel), ratio between old members and total members in the system (middle panel), and ratio between new groups and active members(bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

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

Table 1.1: Jensen Shannon divergence between group sizes distributions from model (in model we vary affiliation parameter paff) and data.

1.4 Distributions fit

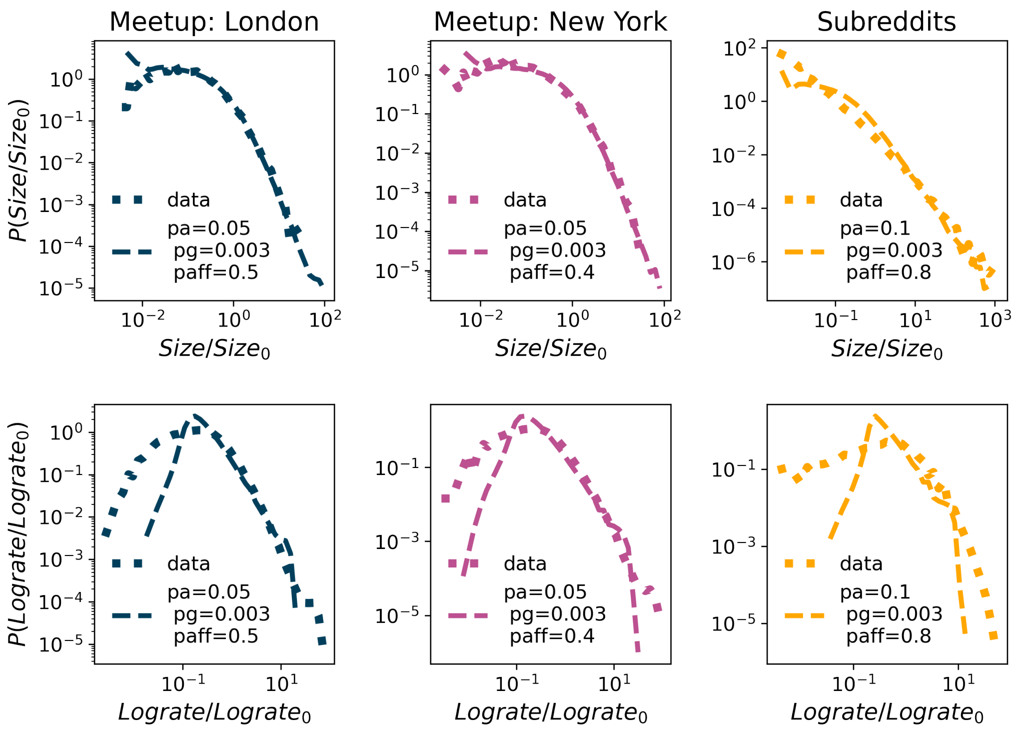

Figure 1.6: The comparison between empirical and simulation distribution for group sizes (top panel) and logrates (bottom panel).

Table 1.2: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **groups sizes** of Meetup groups in London, New York and in Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

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

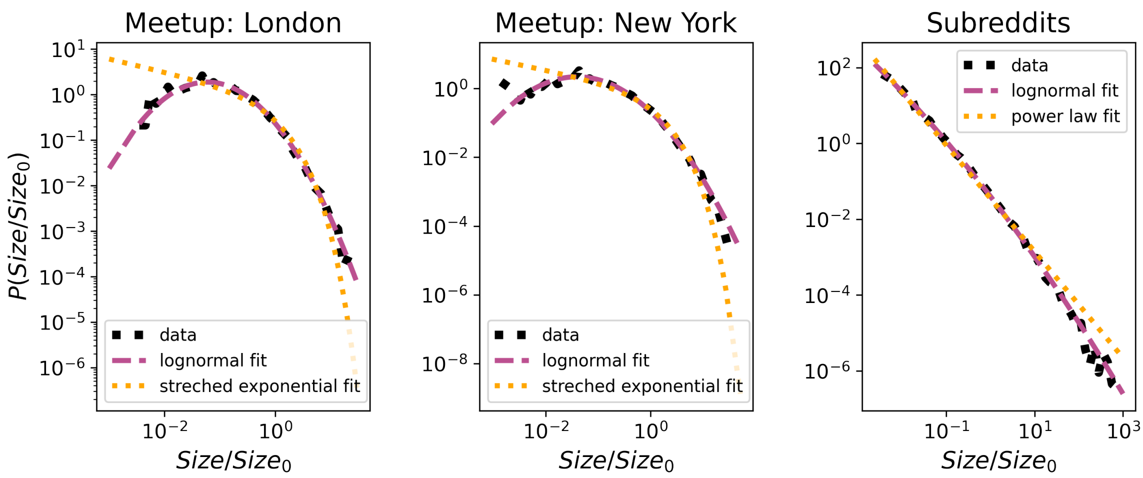

Figure 1.7: The comparison between log-normal and stretched exponential fit to London and NY data, and between log-normal and power law for Subreddits. The parameters for log-normal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for Subreddits $\mu = -5.41$ and $\sigma = 3.07$.

Table 1.3: The likelihood ratio R and p-value between different candidates and **lognormal** distribution for fitting the distribution of **simulated group sizes** of Meetup groups in London, New York and Reddit. According to these statistics, the lognormal distribution represents the best fit for all communities.

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

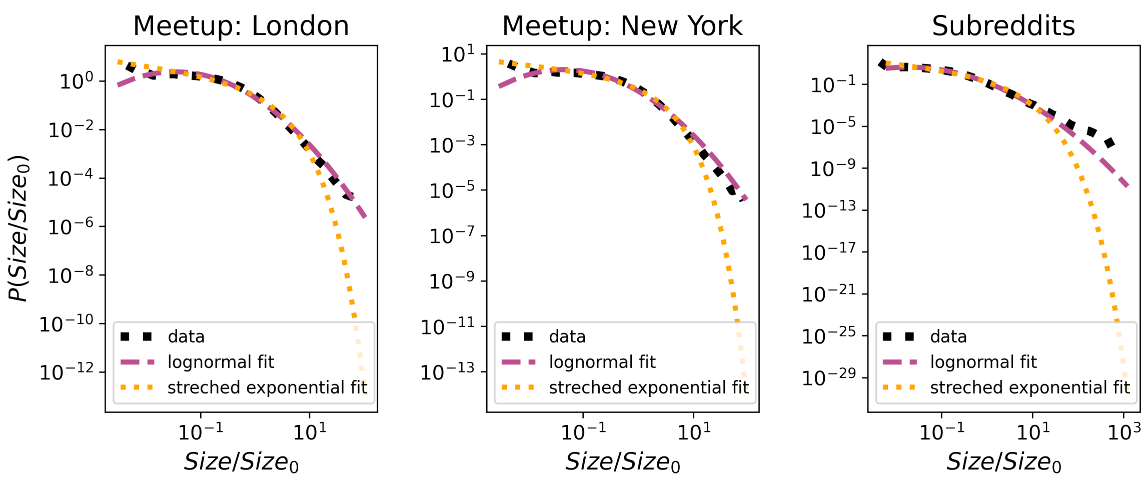

Figure 1.8: The comparison between lognormal and stretched exponential fit to simulated group sizes distributions. The parameters for log-normal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma = 1.53$.

Bibliography

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

[2] Hernan Mondani, Petter Holme, and Fredrik Liljeros. Fat-tailed fluctuations in the size of organizations: the role of social influence. *PLoS One*, 9(7):e100527, 2014.

[3] Dongfeng Fu, Fabio Pammolli, Sergey V Buldyrev, Massimo Riccaboni, Kaushik Matia, Kazuko Yamasaki, and H Eugene Stanley. The growth of business firms: Theoretical framework and empirical evidence. *Proceedings of the National Academy of Sciences*, 102(52):18801–18806, 2005.

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

[5] Jiang-Hai Qian, Qu Chen, Ding-Ding Han, Yu-Gang Ma, and Wen-Qing Shen. Origin of gibrat law in internet: Asymmetric distribution of the correlation. *Physical Review E*, 89(6):062808, 2014.

[6] Sanjay Ram Kairam, Dan J Wang, and Jure Leskovec. The life and death of online groups: Predicting group growth and longevity. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 673–682, 2012.

[7] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1007–1016, 2009.

[8] Marija Mitrović, Georgios Paltoglou, and Bosiljka Tadić. Quantitative analysis of bloggers' collective behavior powered by emotions. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02005, 2011.

[9] Marija Mitrović Dankulov, Roderick Melnik, and Bosiljka Tadić. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Scientific reports*, 5(1):1–10, 2015.

[10] Ana Vranić and Marija Mitrović Dankulov. Growth signals determine the topology of evolving networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(1):013405, 2021.

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

[12] Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 44–54, 2006.