
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

Both datasets have the same structure (*group-id, user-id, timestamp*), where the timestamp was the date when the user joined the group. For each time step, we can calculate the number of new members in each group $N_i(t)$, and the group size $S_i(t)$. The group size at time step t is $S_i(t) = \sum_{k=t_0}^{k=t} N_i(k)$, where t_0 is month when group is created. The group size is increasing in time, as we do

¹<https://www.reddit.com/>

²www.meetup.com

1. The growth of social groups

not have information if the user left the group. Also we calculate the growth rate, as the logarithm of successive sizes $R = \log(S_i(t)/S_i(t-1))$.

Even though Meetup and Reddit are different online platforms, we find some common properties of these systems; see figure 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:

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

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.

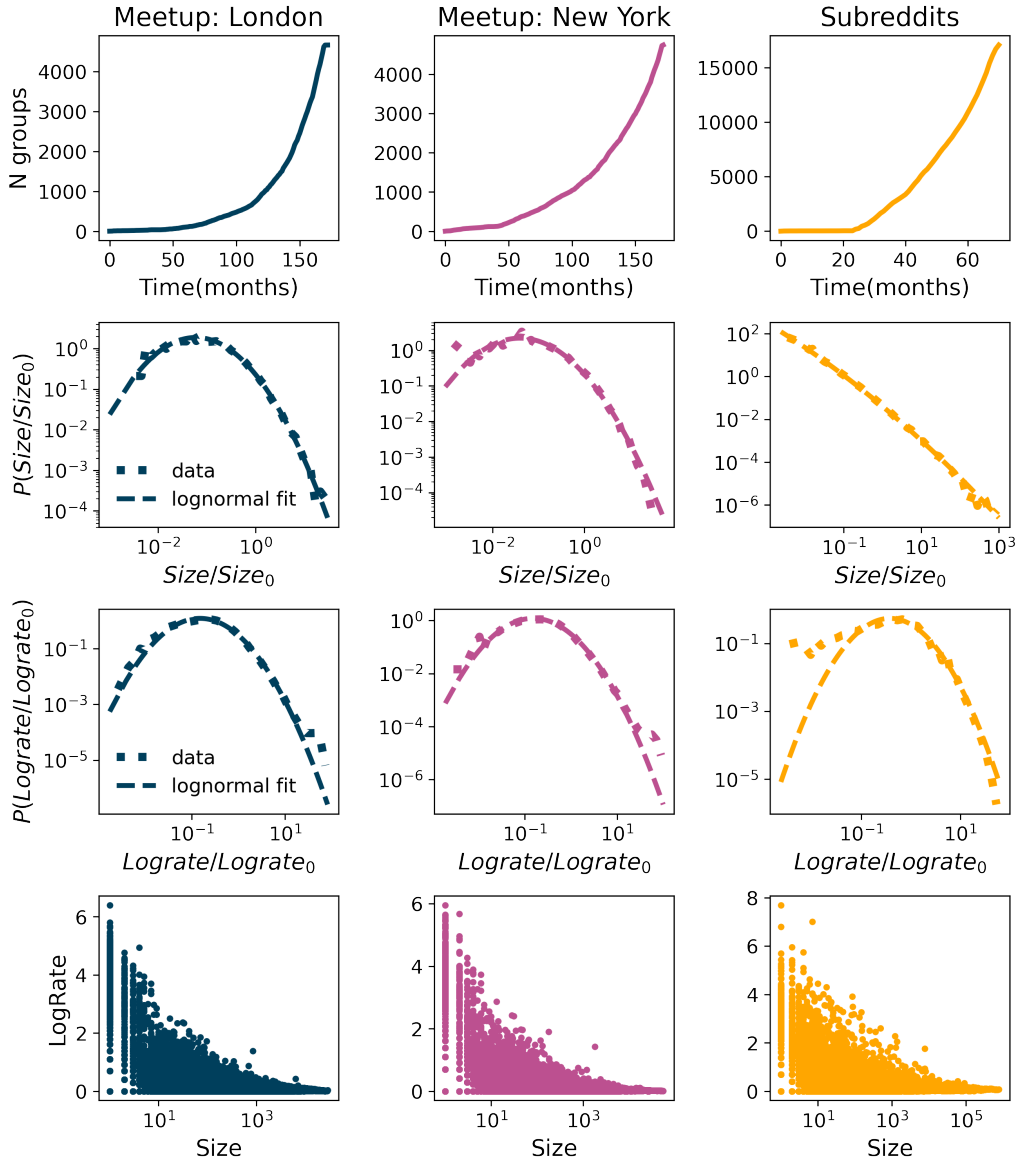


Figure 1.1: The number of groups over time, normalized sizes distribution, normalized log-rates distribution and dependence of log-rates and group sizes for Meetup groups created in London from 08-2002 until 07-2017 that were active in 2017 and subreddits created in the period from 01-2006 to the 12-2011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

The multiplicative process [6] is the simplest way to generate the lognormal distribution. It is also known as Gibrat law and it states that the growth rates are uncorrelated $R = \log \frac{S_t}{S_{t-\Delta t}}$. The distribution of sizes $P(S)$ follow lognormal while the distribution of log rates should follow the normal distribution or, as shown in many studies explained with Laplacian (“tent shaped”) distribution [7], [8]. For Meetup and Reddit, we find that the lognormal distribution fits logrates. Furthermore, logrates depend on the group size. For these reasons, the growth of social groups can not be explained with Gibrat law, [9, 10].

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

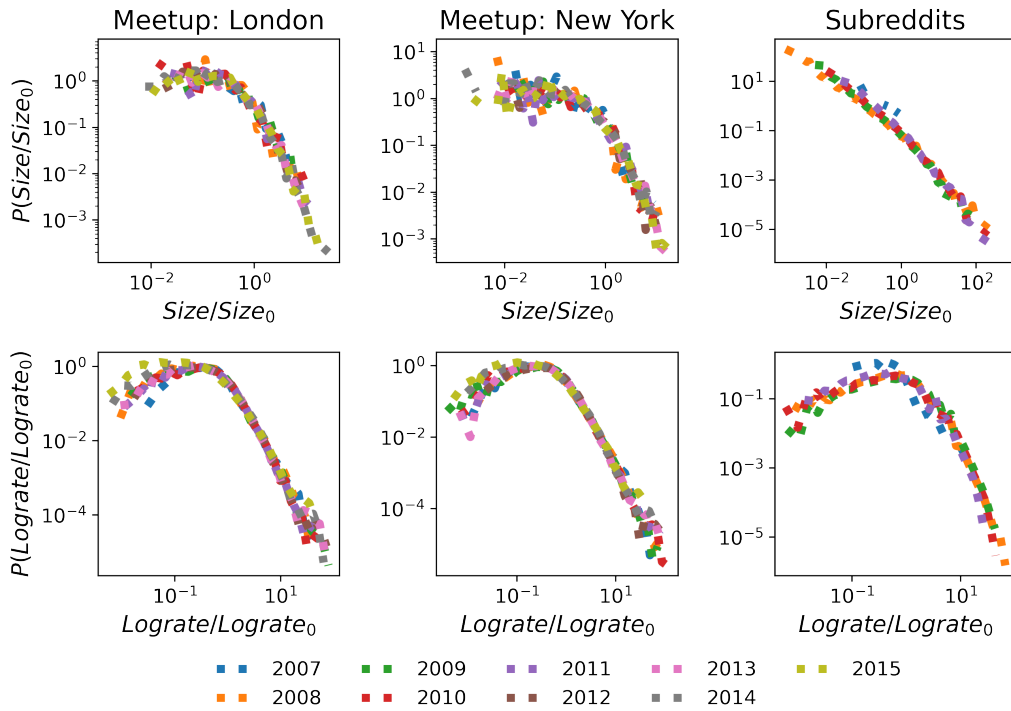


Figure 1.2: The figure shows the groups’ sizes distributions and log-rates distributions. Each distribution collects groups founded in the same year and is normalized with its mean value. The group sizes are at the end of 2017 for meetups and 2011 for subreddits.

1.2 The groups growth model

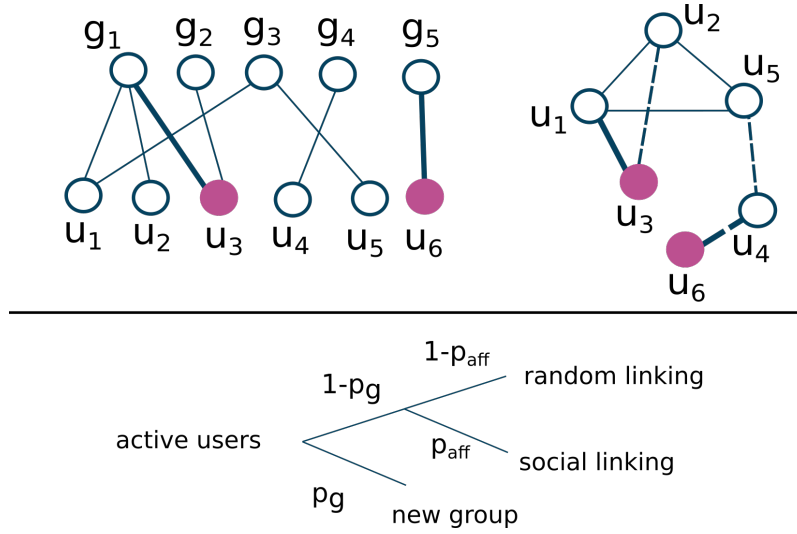


Figure 1.3: The top panel shows bipartite (member-group) and social (member-member) network. Filled nodes are active members, while thick lines are new links in this time step. In the social network dashed lines show that members are friends but still do not share same groups. The lower panel shows model schema. **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 .

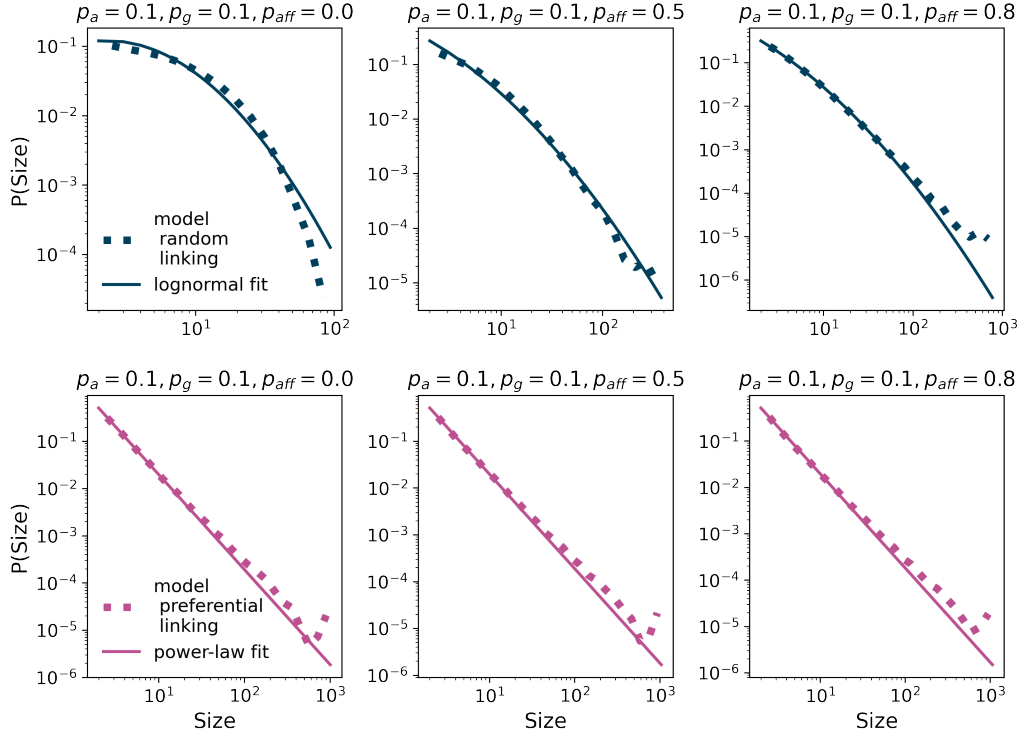


Figure 1.4: Groups sizes distributions for groups model, where at each time step the constant number of users arrive, $N = 30$ and old users are active with probability $p_a = 0.1$. Active users make new groups with probability $p_g = 0.1$, while we vary affiliation parameter p_{aff} . With probability, $1 - p_{aff}$, 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

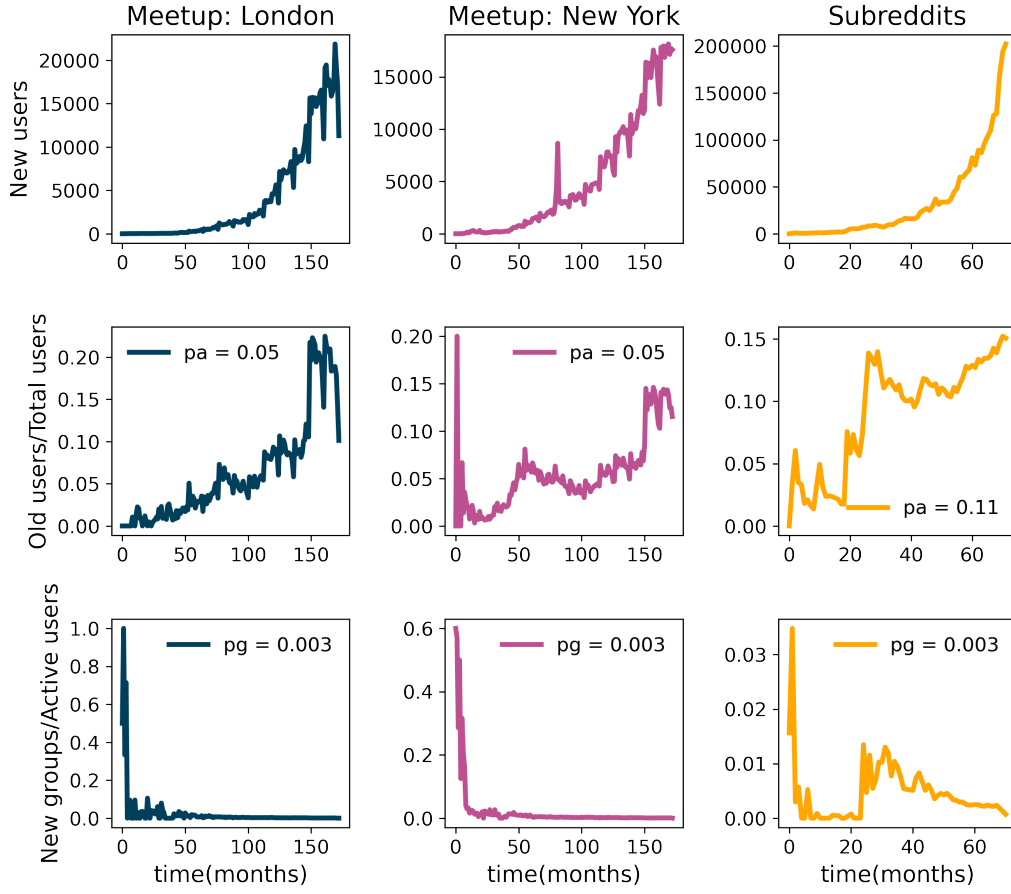


Figure 1.5: The time series of number of new members (top panel), ratio between old members and total members in the system (middle panel), and ratio between new groups and active members (bottom panel) for Meetup groups in London, Meetup groups in New York, and subreddits.

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 p_{aff}) and data.

1.4 Distributions fit

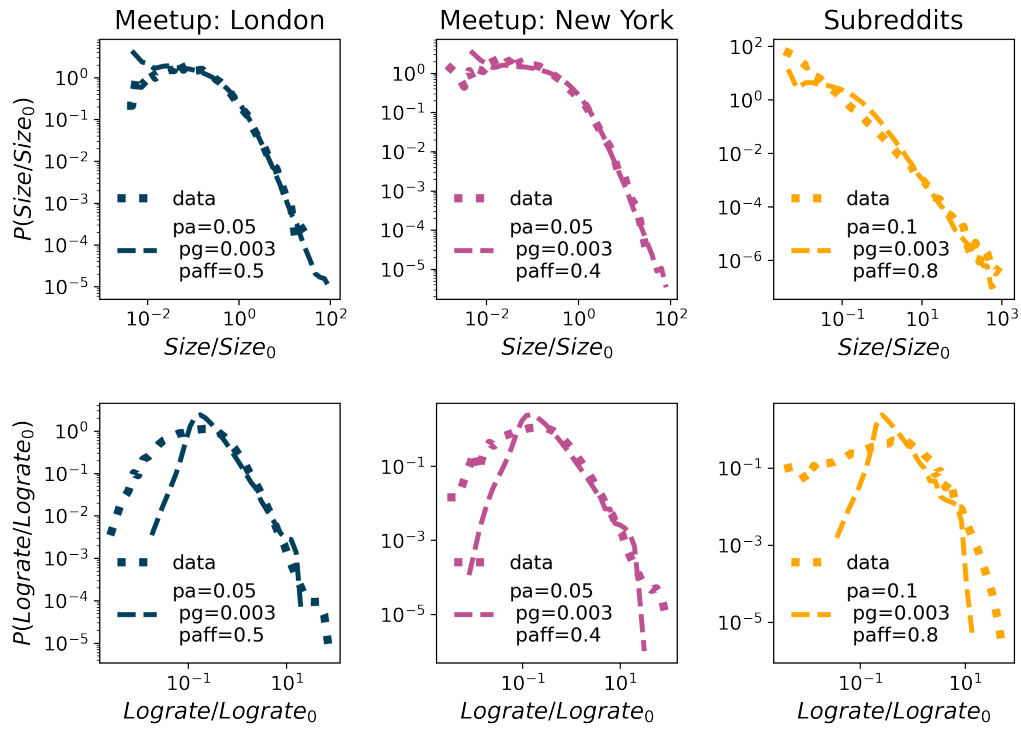


Figure 1.6: The comparison between empirical and simulation distribution for group sizes (top panel) and logrates (bottom panel).

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

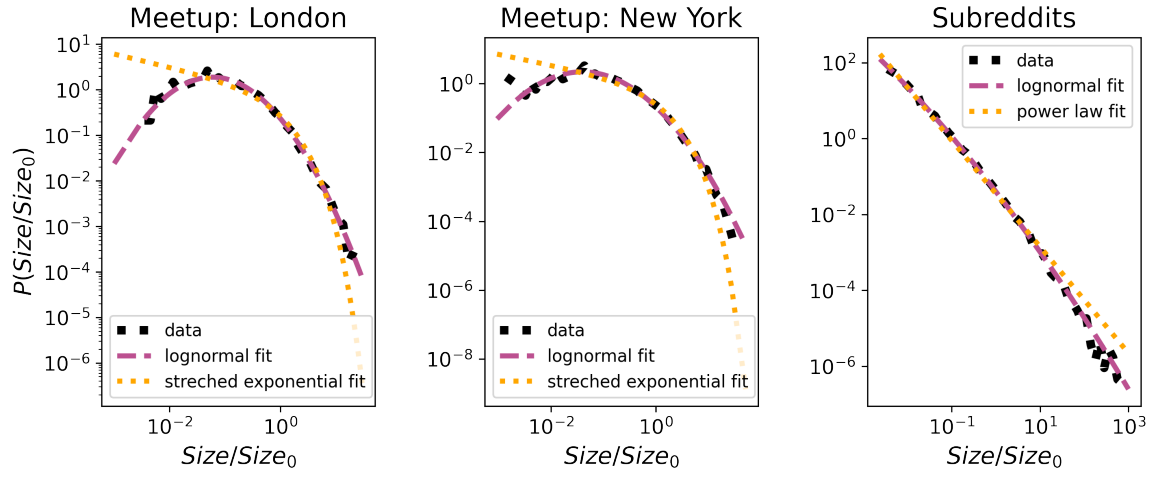


Figure 1.7: The comparison between log-normal and stretched exponential fit to London and NY data, and between log-normal and power law for Subreddits. The parameters for log-normal fits are 1) for city London $\mu = -0.93$ and $\sigma = 1.38$, 2) for city NY $\mu = -0.99$ and $\sigma = 1.49$, 3) for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

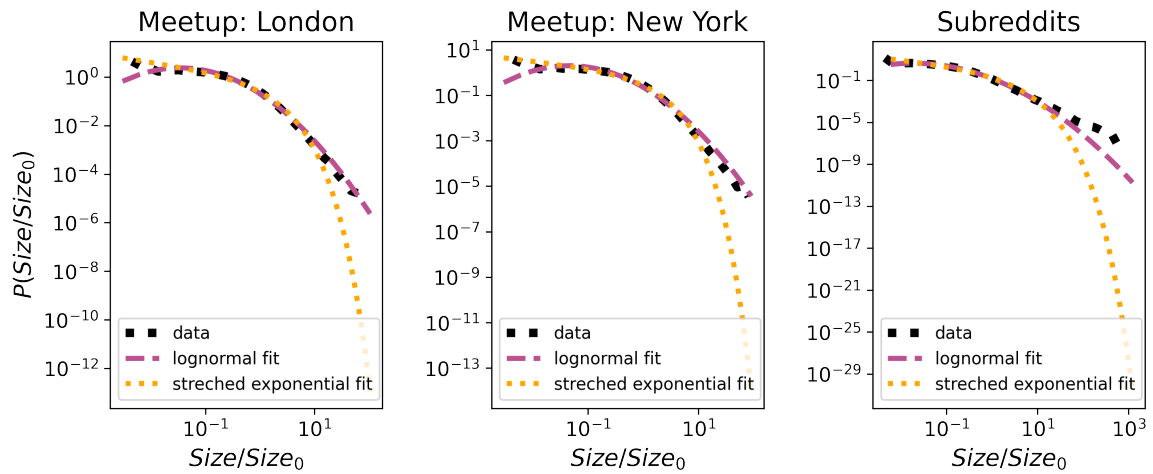


Figure 1.8: The comparison between lognormal and stretched exponential fit to simulated group sizes distributions. The parameters for log-normal fits are 1) for city London $\mu = -0.97$ and $\sigma = 1.43$, 2) for city NY $\mu = -0.84$ and $\sigma = 1.38$, 3) for Subreddits $\mu = -1.63$ and $\sigma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

Bibliography

- [1] Sergey N Dorogovtsev and José FF Mendes. Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E*, 63(5):056125, 2001.
- [2] Jelena Smiljanić and Marija Mitrović Dankulov. Associative nature of event participation dynamics: A network theory approach. *PloS one*, 12(2):e0171565, 2017.
- [3] Milovan Suvakov, Marija Mitrovic, Vladimir Gligorijevic, and Bosiljka Tadic. How the online social networks are used: dialogues-based structure of myspace. *Journal of The Royal Society Interface*, 10(79):20120819, 2013.
- [4] Tiago A Schieber, Laura Carpi, Albert Díaz-Guilera, Panos M Pardalos, Cristina Masoller, and Martín G Ravetti. Quantification of network structural dissimilarities. *Nature communications*, 8(1):1–10, 2017.
- [5] Hernán A Makse, Shlomo Havlin, Moshe Schwartz, and H Eugene Stanley. Method for generating long-range correlations for large systems. *Physical Review E*, 53(5):5445, 1996.
- [6] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.
- [7] 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.
- [8] 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.
- [9] Gerald F Frasco, Jie Sun, Hernán D Rozenfeld, and Daniel Ben-Avraham. Spatially distributed social complex networks. *Physical Review X*, 4(1):011008, 2014.
- [10] 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.