Assignment 2: Analyzing the Structure of a Social Network

Research Report

Mai Ngoc Nong

Department of Network and Data Science

DNDS6014: Introduction to Computational Social Science

Prof. Mark Wittek

November 9, 2024

1. Introduction

The LastFM Asia Social Network dataset, which consists of user interactions and musical preferences, provides a unique opportunity to conduct structural analysis of a typical social network. This study primarily focuses on network structural analysis by assessing the degree distribution to determine whether it exhibits traits of a scale-free network where there are a few high-degree nodes and most have few neighbors.

2. Data Preprocessing

The data preprocessing starts with loading the edge list to create the network using Python's 'networkx' library. Each node in the dataset represents a user and each edge is formed by the mutual connections between a pair of users on the platform. The feature associated with each node represents the list of artists that the corresponding user follows. The dataset also comes with a target table containing users' self-identified locations, which will be used to add more information for the network visualization.

3. Data Analysis

The network is visualized using 'networkx' and 'matplotlib'. To add more layers of information to the network, the nodes' sizes are scaled by a factor of 10 with degree, or number of neighbors, and the nodes are color-coded using the target values (i.e., user's location)

To explore the scale-free nature of the network, the degree distribution along with the degree probability density function of the network is first visualized on a log-log scale to quickly inspect the shape and skewness of the distribution. The degree distribution is then fit to a power-law distribution using the 'powerlaw' library to obtain the alpha, which is the exponent that defines the heaviness of the distribution's tail. The fit result also returns a Kolmogorov-Smirnov (KS) statistic to measure the maximum difference between the cumulative distribution of the network

degrees and that of the theoretical power-law model with the estimated parameters. Since 'powerlaw' does not return a p-value off the shelf to assess the goodness-of-fit, a bootstrap test was then performed to examine if the observed KS statistic is within the 95% confidence interval of the KS statistics generated by fitting the power-law distribution to the bootstrap samples of the network's degrees.

It should also be noted that other heavy-tailed distributions, such as the log-normal, exponential, and truncated power-law distributions, can also provide a better fit for the data and indicate scale-free behavior. Therefore, these distributions will be compared against the power-law distribution using the loglikelihood ratio to see which distribution is preferred over the power-law distribution.

4. Results and Discussion

4.1. Network Visualization

The constructed network contains 7624 nodes corresponding to users and 27806 edges (Figure 1). The network has an average clustering coefficient $C = 0.2194$ and an average shortest path length of $L = 5.2322$. Figure 1 provides a quick snapshot of the LastFM Asia social network. Overall, the 3 locations with the highest number of users are: 0, 10, 17. Visual inspection shows no extremely high-degree nodes as most users have few mutual connections.

4.2. Degree distribution

Figure 2 displays the distribution and the probability density function of the network degrees. Initial fitting to a power-law distribution obtained an exponent $\alpha = 3.3263$ and the Kolmogorov-Smirnov (KS) statistic $D = 0.0503$, which is within the 95% confidence interval (0.0362, 0.0612) of the bootstrap KS statistics (shown in Figure 3), suggesting that the power-law distribution provides a reasonable fit for the network's degrees.

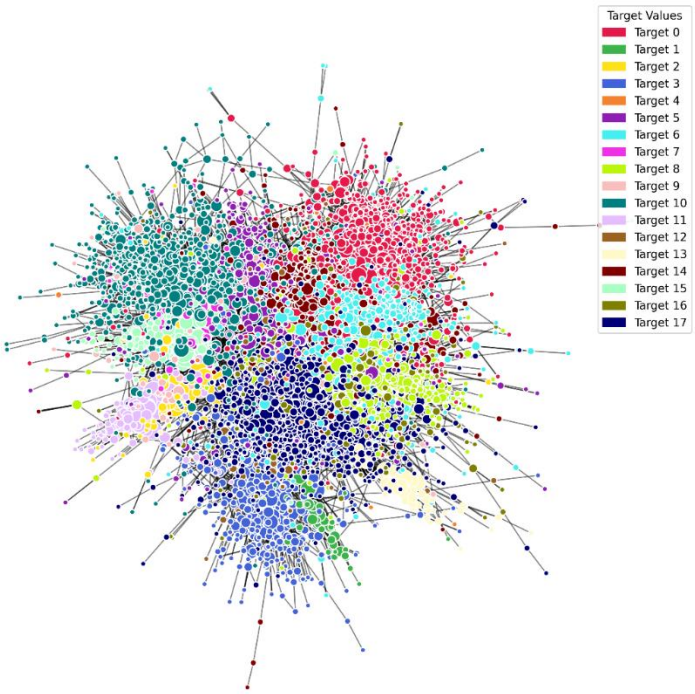

Fig. 1. LastFM Asia Social Network (node sizes corresponding to degree, node color corresponding to self-identified location of user)

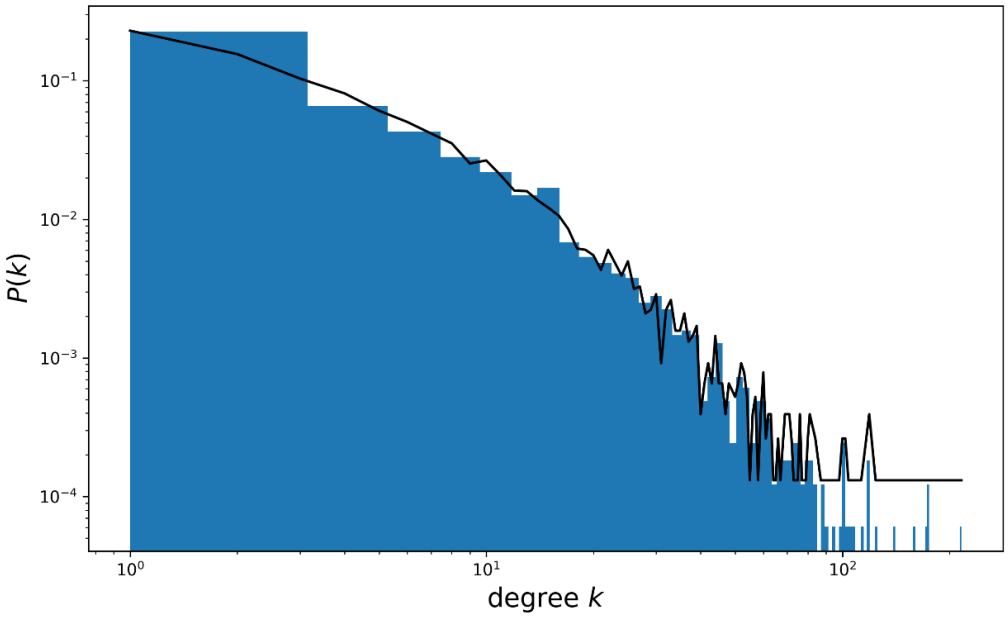

Fig. 2. Histogram and degree probability density function of the network degrees on a log-log scale

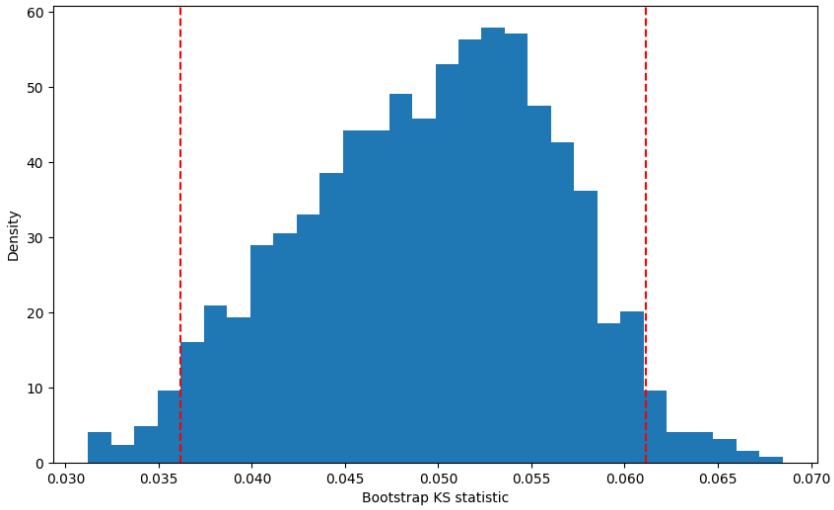

Fig. 3. Distribution of bootstrap samples' KS statistic compared to a power-law distribution (red dotted lines correspond to 95% confidence interval)

Comparison results between each pair of candidate distributions (Table 1.) indicate that the truncated power-law distribution provides the best fit out of the four distributions. This result suggests that the network is scale-free but with some natural constraints or boundaries limiting the degree distribution, like network size, or the platform's systemic limitation on the number of connections a user can have, which prevents extremely high-degree nodes. The probability density function of the network's degrees in comparison with those of the candidate distributions is visualized in Figure 4.

Distribution 1	**Distribution 2**	**Loglikelihood Ratio**	**p-value**
power_law	truncated_power_law	-2.015032	0.044696
power_law	lognormal	-1.498698	0.276950
power_law	exponential	9.395554	0.144588
truncated_power_law	lognormal	0.516334	0.041487
truncated_power_law	exponential	11.41059	0.024377
lognormal	exponential	10.89425	0.033522

Table 1. Loglikelihood ratio between pairs of candidate distributions. Positive loglikelihood ratio indicates distribution 1 is preferred, whereas negative loglikelihood ratio indicates distribution 2 is preferred.

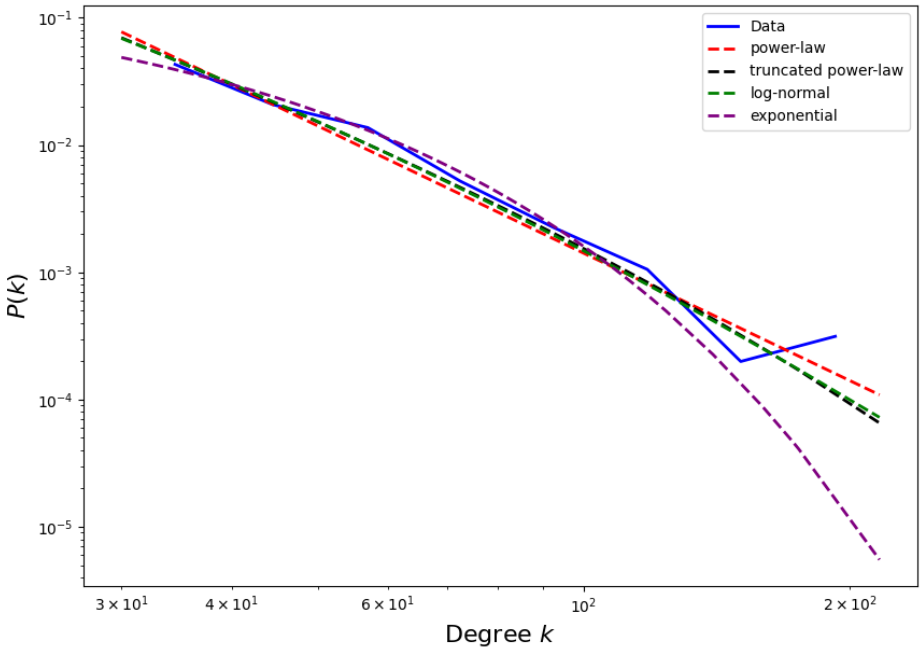

Fig. 4. Probability density function of the network's degrees (blue line) in comparison with candidate distributions (dotted lines)

5. Conclusion

The analysis of the LastFM Asia Social Network's structure by examining the degree distribution has confirmed its scale-free nature. Initial inspection of the degrees' histogram and probability density function as well as comparison to a power-law distribution supported the hypothesis that most users in the network only have a few mutual connections. Furthermore, it was revealed through subsequent comparisons of fit among heavy-tailed distributions that the truncated power-law distribution provides a better fit for the distribution of degrees, suggesting that scale-free behavior only exists up to a certain threshold, and the probability of higher degree nodes decays more substantially than in a pure power-law distribution. While this study has only examined the overall network structure through mutual connections among users, subsequent studies exploring modules within the network can follow a similar pattern to compare how scale-

free behavior differs from one community in the network to another, taking into account other factors such as community size, homogeneity in users' location or musical preferences, and so on.

References

Alstott J., Bullmore E., & Plenz D. (2013). powerlaw: A Python Package for Analysis of Heavy-Tailed Distributions. PLOS ONE 9*(4): e85777.* https://doi.org/10.48550/arXiv.1305.0215

Hagberg A. A., Schult D. A., & Swart P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (pp. 11–15). Pasadena, CA: SciPy2008.