Analysis of the Facebook Network

Miguel Lourenço Farinha (86798)

Instituto Superior Técnico
Master in Mathematics and Applications
Complex Network Science
1^{st} Semester- 2020/2021

Abstract

This is a study which aimed to analyse the social network of Facebook. Several global metrics, centrality measures and degree correlation metrics were applied. Furthermore, link prediction algorithms were explored as preparation for the second Project. The dataset was analysed resorting to NetworkX.

1. Introduction

The Facebook network was obtained from the Stanford Large Network Dataset Collection [4]. This dataset consists of 'friends lists' from Facebook. The dataset includes node features (profiles), circles, and ego networks [4]. *Ego-centered networks* are networks surrounding one particular individual, designated as *ego*, meaning, usually, the individual surveyed and his or her immediate contacts, designated as *alters* [5]. This dataset is an undirected graph with a single strongly connected component. A preliminary analysis of the dataset allowed the compilation of Table 1.

Number of nodes	4039
Number of edges	88234
Average degree	43.69
Average clustering coefficient	0.6055
Maximum degree	1045
Minimum degree	1

Table 1: Dataset statistics.

Firstly, in order to characterize the network's global structure the degree distribution, the clustering coefficient and the average path length were computed. Secondly, a variety of measures that captured particular features of the network topology were considered, namely, the degree centrality, the betweeness centrality, the closeness centrality and the eigenvector centrality. Further measures, not mentioned in this report, were applied. The degree correlations of the network were also studied through the application of some measures. The values obtained for the computed metrics were discussed and their influence on the network's topology was examined. To further study the network and as preparation for the second project, link prediction techniques were explored.

2. Results & Discussion

The results and its respective discussion were carried out simultaneously throughout this report.

2.1. Global measures

The aim of this section was to acquire an overview of the global structure and topology of the network.

Firstly, the degree distribution of the network was computed. The degree distribution corresponds to the probability distribution of the degrees of all the nodes in the network. This measure can be computed by applying the following expression:

$$p_k = \frac{n_k}{N} \tag{1}$$

where n_k is the number of nodes with degree k and N is the total number of nodes. Therefore p_k can be thought of as the probability that a randomly chosen node in the network has degree k. By plotting the degree distribution as a function of the degree it was possible to observe that most of the nodes had a low degree. However, the computed plot showed a significant "tail" to the distribution which corresponded to the nodes with substantially higher degree, *i.e.*, the hubs of the network. Therefore, the degree distribution was right-skewed. The previously mentioned plot was then obtained using logarithmic scales which yielded the graph depicted in Figure 1.

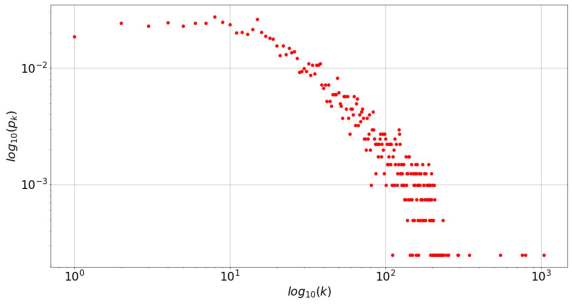

Figure 1: Degree distribution of the Facebook network.

It would be reasonable to expect that the degree distribution would follow a power law degree distribution. However, the leftmost region of the graph

of Figure 1 showed a low-degree saturation (flattening of the curve) which indicated that the number of nodes with small degree was fewer than expected. Furthermore, the rightmost region of the graph showed a high-degree saturation characterized by noisy points and a very steep drop of p_k, due to the fact that the number of nodes with a large degree was reduced. To better visualize the degree distribution the cumulative distribution function was computed. In Figure 2 the complementary cumulative distribution function (CCDF) and the power law distribution function (which best fitted the data) were plotted.

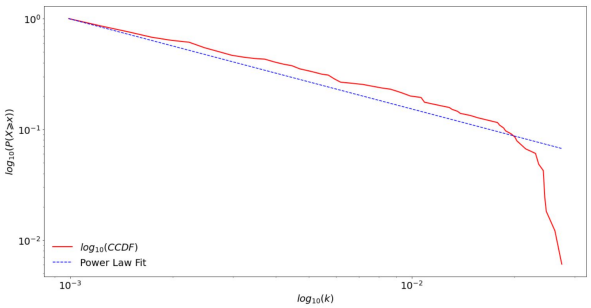

Figure 2: Complementary cumulative distribution function for the degrees of the nodes and power law fit.

Several conclusions were withdrawn from the observation of Figure 2. Firstly, it was verified that the CCDF for the degrees of the nodes did not follow a Poisson distribution. Thus, the degree distribution of the network was distinct from the one obtained for a theoretical random network. Secondly, for large values of k, the distribution failed to follow a power law distribution as evidenced by the observed cutoff. Since the network was finite, according to [6], there was a maximum number of connections a node could have which explained the observed cutoff. Finally, considering that the distribution p_k follows, to some extent, a power law, then so does the cumulative distribution function P_k, but with an exponent $\gamma - 1$ that is 1 less than the original exponent [5]. The computed value for γ was 2.81 with a standard deviation of 0.06. Despite the fact that the computed γ was in the range of common values for the parameter for scale-free networks, it would not have been appropriate to classify this network as scale-free. Instead following the classification of graphs proposed by [2] it would be appropriate to state that this network, probably, should be classified as a broad-scale network.

The clustering coefficient is a measure of the degree to which nodes in a graph tend to cluster together. To calculate this measure for an undirected graph one resorts to the equation below:

$$C_i = \frac{e_i}{\frac{k_i(k_i-1)}{2}} \tag{2}$$

where e_i is the number of edges among the neighbours of node i and k_i is the degree of node i. The average clustering coefficient computed for the network was 0.6055. The plot of the average clustering coefficient as a function of the degree was depicted in Figure 3.

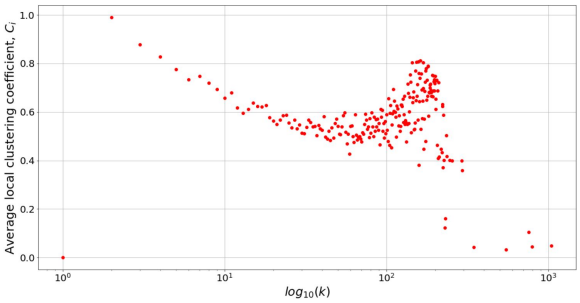

Figure 3: Local clustering coefficient as a function of degree.

An analysis of Figure 3 corroborated the computed value of 0.6055 for the average clustering coefficient. Given the obtained values for C_i, which measures the network's local link density, it was possible to conclude that on average the neighbourhood of each node of the network was densely connected which was an expected property of a social network. According to [5], nodes with higher degree tend to have, on average, a lower clustering coefficient. Local clustering may be used as probe for the detection of *structural holes* on the network which are links among neighbours of a node that one would expect to find on a social network. In *ego networks*, *structural holes* are of interest for the hub of the network whose neighbours lack connections, because they will give the hub power over the information flow between those neighbours [5]. The local clustering coefficient measures how influential a node is in this sense, taking lower values the more *structural holes* there are in the network around the node. Inspection of Figure 3 revealed that this was the case with node 107 (node with largest degree) which had a clustering coefficient of 0.049. Thus, this node possessed a considerable control over the flow of information within the network. Moreover, the slope of the plot in Figure 3 followed approximately a negative pattern which led to the conclusion that the network had an hierarchical structure.

The average path length is defined as the average number of steps along the shortest paths for all possible pairs of network nodes and is a measure of the efficiency of information transport on a

network. This measure can be computed with the following expression:

$$< L >= \frac{1}{N(N-1)} \times \sum_{ik} L_{ik} \quad (3)$$

where N is the total number of nodes and L_{ik} is the length of the shortest path between nodes i and k. The obtained value for this measure was 3.69. Therefore, one could infer that the network possessed the *small-world* property as proposed by [7]. Hence, information spreading between nodes was extremely efficient and fast. According to [5], mathematical models of networks suggest that path lengths in networks should typically scale as $log(N)$ with the number N of network nodes, and should therefore tend to remain small even for large networks because the logarithm is a slowly growing function of its argument. Taking the logarithm of the number of nodes of the network corroborates this assumption since $log(4039) \approx 3.60 \sim 3.69$.

2.2. Centrality measures

Centrality measures aim to identify the most important nodes within a network and capture particular features of its topology. This metrics are particularly important, for example, in the field of social network analysis. As discussed in the introduction the Facebook network was an *ego-centered network*. Moreover, its graph had a single strongly connected component which meant that the network's hub was connected to all other nodes in the network, whether that connection was direct or indirect. Therefore, it would be expected that the most important node within the network was the *ego* node which in this case corresponded to node 107. To assess the veracity of the latter statement several centrality measures were applied.

Firstly, the degree centrality was measured for all nodes. The highest result was obtained for node 107 with a value of 0.258, which was expected beforehand due to the node's high degree. This result seemed consistent with the definition of an *ego-centered network*, since the *ego* node, which had the largest number of connections, probably had more influence, more access to information and more prestige than those who had fewer connections. A disadvantage of degree centrality is that it only captures the immediate connections a given node has, rather than indirect connections to all other nodes. In other words, node 107 was connected to a large number of nodes, but such nodes might have been rather disconnected from the network as a whole. Consequently, node 107 could be quite central, but only in a local neighbourhood. Therefore, the closeness centrality for all nodes was computed. Once again, node 107 yielded the highest value. This result was coherent with the fact that node 107 has the highest potential to spread information throughout the network due to its centrality. Thirdly, the betweeness centrality (which differs from the other centrality measures in being not principally a measure of how well-connected a node is) for all nodes was obtained. Similarly, the highest computed value was obtained for node 107. This measure quantifies the number of times a node acts as a bridge along the shortest path between two other nodes and, as expected, node 107 had a considerable influence and control over the flow of information passing between all other nodes. Lastly, the eigenvector centrality measure was computed. This metric assigns relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. In contrast with the previous measures, node 107 did not have the highest eigenvector centrality. This was probably due to the fact that this node was connected to a large number of other nodes which were not important in the network. This interpretation is consistent with real life social circles which have an *ego*, *i.e.*, a person which knows everyone regardless of their importance and connections. The pagerank and the harmonic centrality were also computed, yielding results which were in line with the previous results and interpretations.

2.3. Degree correlations

In social networks hubs tend to date each other. Therefore, it was of interest to study if this pattern did in fact happen in this network. The degree correlations between the nodes and their implications in the network's topology were studied. Nodes tend to attach to other nodes that are similar in some way, in other words, the network is said to be assortative. In this case, the similarity measure is the degree of a node [3]. Contrarily to the idea that in social networks hubs tend to date each other, in the case of the studied network it would be reasonable not to expect such a behaviour. Since the network is centered around the *ego*, it would make sense for this node to be connected with nodes whose degrees vary through a wide range of values instead of just connecting with the other hubs. If the *ego* node would happen to just connect with other hubs then its influence and power over the network would be impaired, since it would be dependent on the other hubs to access and control information. Therefore, the network was not expected to have an assortative behaviour. In order to corroborate the last assumption the Pearson correlation coefficient of the network was computed, yield-

ing a value of 0.06. The obtained value was in agreement with the hypothesis previously stated. Moreover, since the Pearson correlation coefficient was close to zero the network was neither assortative nor disassortative. Therefore, small and high-degree nodes connect to each other which led to the conclusion that the network exhibited the behaviour of a neutral network. Nevertheless, the network's neutral behaviour did not happen randomly. Instead, the nodes with the largest number of degrees selectively chose nodes, regardless of its degrees, that would propel their influence and control over the network thus becoming *egos*. Degree correlations capture the relationship between the degrees of nodes that link to each other [1]. One way to quantify their magnitude is to measure for each node i the average degree of its neighbours. The degree correlation function calculates for all nodes with degree k:

$$k_{nn}(k) = \sum_{k'} k' P(k'|k) \tag{4}$$

where $P(k'|k)$ is the conditional probability that following a link of a k-degree node we reach a degree-k' node. Therefore $k_{nn}(k)$ is the average degree of the neighbours of all degree-k nodes [3].To quantify degree correlations we inspect the dependence of $k_{nn}(k)$ on k. Figure 4 presented such dependence on a linear-log scale.

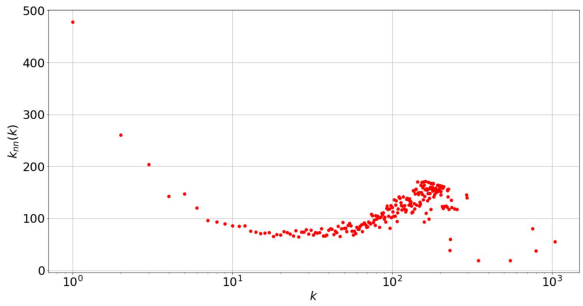

Figure 4: Degree correlation function.

The plot exhibited, approximately, an horizontal pattern. Therefore, such behaviour of the degree correlation function indicated a lack of degree correlations, which is in line with the value obtained for the Pearson correlation coefficient. Moreover, one may state to some extent that $k_{nn}(k)$ is independent of k.

3. Link Prediction

Due to the close relation with Machine Learning, link prediction techniques were explored. Firstly, some simple metrics and tools were explored such as the Jaccard coefficient, preferential attachment and common neighbours. For example, a simple analysis of the results led to the conclusion that the preferential attachment method prioritized the connection between nodes with higher degree as opposed to the Jaccard coefficient. Thereafter, a Logistic Regression model to predict future links was implemented. However, due to computational issues it was not possible to assemble the complete model. Future work should involve testing further link prediction methods. This subject may be the basis of the second project of the Complex Network Science course.

4. Conclusions

The Facebook network had a topology and structure highly determined by the fact that this network was centered around the *ego* node as demonstrated by all the computed measures. In conclusion, the Facebook network should be classified as a broad-scale network with an hierarchical structure. Its most important node was the *ego* of the network and the connections between nodes followed a neutral behaviour. Link prediction is a subject with many relevant applications which should be the object of further study.

References

[1] A.L. Barabási. *Network Science*. 2016.

[2] L.A.N. Amaral, A. Scala, M. Barthélémy and H. E. Stanley. Classes of small-world networks.

[3] Francisco Correia dos Santos. Complex network science lecture notes 2020/2021.

[4] J. McAuley and J. Leskovec. Learning to discover social circles in ego networks. NIPS, 2012.

[5] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press Inc., New York, 2010.

[6] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin. Size-dependent degree distribution of a scale-free growing network. *Physical Review E*, (63), 2001.

[7] J. Travers and S. Milgram. An Experimental Study of the Small World Problem. *American Sociological Association*, 32(4):425–443, 1969.