Project Report - Complex Networks Science

Group 90: 82033 - Ana Cardoso
82468 - Filipe Azevedo
82517 - Martim Zanatti

2018/2019

1 Introduction

In this assignment — from the available datasets options — we chose to analyze one that we thought was interesting and big enough to explore and analyze the network characteristics. This work was made considering the second suggestion option, given that our network it's not too big and we have implemented some code, using python libraries, more precisely Networkx [1].

Our dataset consists of information about 320 YouTube videos crawled and it's related videos from the Video Tops on February 22nd of 2007.

Each instance is composed by a unique ID (11-digits), and the video metadata: uploader, date when it was added, category, length, user rating, number of views, ratings and comments, and a list of "related videos".

We built a graph where each node is a YouTube video and it has an edge to another node if that video is related to it. We assumed that the graph was undirected, which made it possible to apply several metrics, and since if a video is related to the other, then the other video is also related to it.

List of some metrics implemented to analyze the graph:

- Average Degree;
- Cumulative Average Degree Distribution;
- Clustering Coefficient;
- Average Path Length;
- Number of Cliques;
- Different measures of node Centrality;
- Page Rank.

[1] `https://networkx.github.io/`

2 Gathered Results

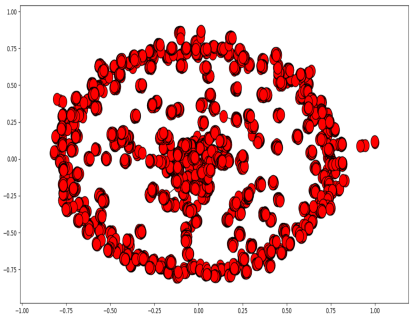


Figure 1: Network architecture.

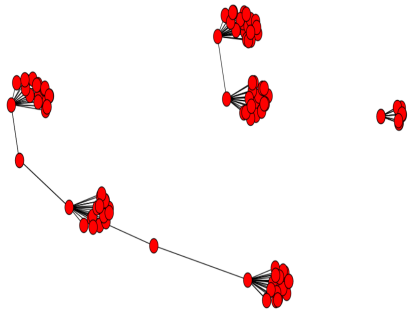


Figure 2: Part of network architecture with more detail.

After parsing the dataset from the text file to the network graph with the necessary information, our network has 5249 nodes and 6045 edges (considering the recommended videos of each instance), as shown in figure 1 and a close up is shown in figure 2.

As can be seen in figure 3, the cumulative degree distribution does not follow a power law. This means that our network is not a Scale Free Network.

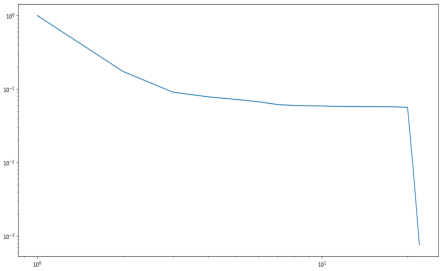
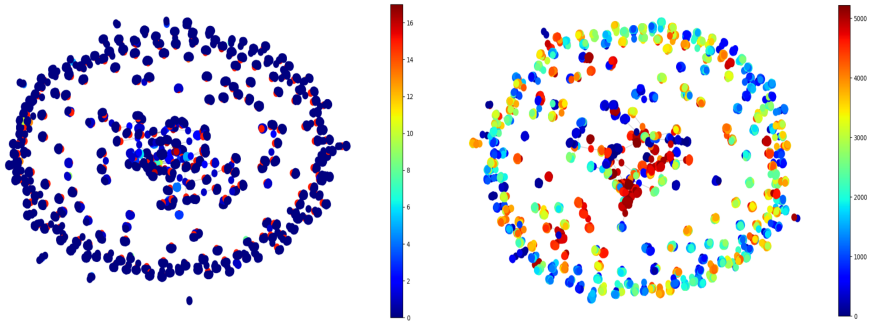


Figure 3: Cumulative Function of Degree Distribution in Log Log Scale.

3 Discussion of the Results

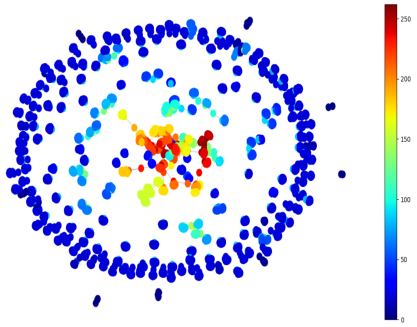
The graph created by the dataset we chose is composed by a core of nodes which correspond to the nodes that have a higher degree, meaning that the nodes that make the core of the graph correspond to videos that are more recommended by other videos. This happened, in part, because there are a lot of videos that were never expanded in the crawl, meaning that we have no information available on them, aside from their ID's and that it is related to the video that it was recommended by.

This can be seen in figure 1: there's a dense core of nodes and a continuous border.

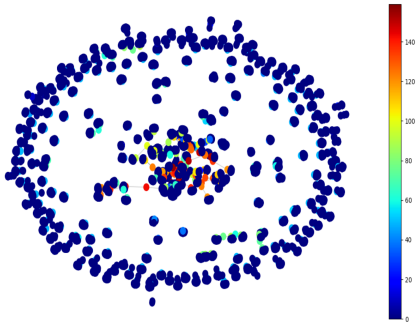


(a) Degree Centrality

(b) Eigenvector Centrality



(c) Closeness Centrality



(d) Betweenness Centrality

Figure 4: Representation of the four centrality measures, nodes colored according to the value of the metric. How to read color scale for each graph: Blue represents the lower value, up to Red, that represents the higher value.

We also came to this conclusion by analyzing figure 4a which gives us the Degree Centrality, or the number of neighbours each node has. In this figure, we can see that most nodes with a lower degree (nodes in blue) are in the border and most nodes with a higher degree (nodes in red) are found in the core. It was expected that this distribution was even, since, as was said before, there are nodes that weren't expanded and therefore have very little neighbours. Despite this we thought that this graph distribution greatly resembles the real distribution of all Youtube videos: there are very little videos that are recommended by a lot

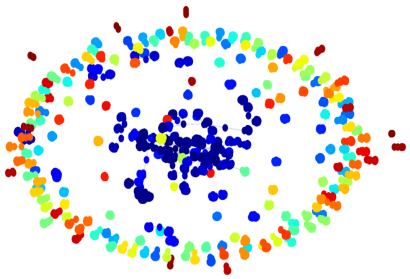


Figure 5: Network after applying Louvain Modularity to detect videos from the same community. Nodes with the same color belong to the same community.

of other videos and a lot of videos that are recommended by very few videos.

We then used the Louvain Modularity to detect communities of videos and through this we were able to see that most neighbours of a node belonged to the same video category of that node. This was something we found to be extremely similar when compared to the real layout of the Youtube graph. There is a core where there are the most important videos, figure 4b, usually the ones with most the views, and through these videos we can get to other videos by the recommendations of less important videos of the same category. It's a network that expands from the most important videos to the least important videos.

4 Code

We submitted the developed code in fénix platform, although we used the Python (3.) library *Networkx* (version 2.2) for performing the computation of some typical measures and the *PowerLaw* library, we had to implement additional code to parse the dataset file (.txt) to a graph, to manipulate the existing functions of these libraries and the visualizations of the graph measures retrieved.

To draw the graph with the visualizations of the centrality measures intended, it is necessary to change one attribute in the main.py file.

The instructions of how to run the developed code are in more detail on the file README.md.

In the terminal shell it will appear a report of the network characteristics, between these informations are the number of nodes, the number of edges, the average degree of the network, the number of triangles and cliques, the power-law gamma of the cumulative degree distribution, the clustering coefficient for each node, the average clustering coefficient, the average path length (for each subgraph of the network), the top 20 of the PageRank values.

Among that, it will appear some graphic drawings, and to make it possible to continue the code to run it is necessary to close those windows.

5 References

- Dataset for "Statistics and Social Network of YouTube Videos" - `http://netsg.cs.sfu.ca/youtubedata/`
- Slides of theoretical lessons - `https://fenix.tecnico.ulisboa.pt/disciplinas/CRC/2018-2019/1-semestre/classes-reading-list-and-supplemental-material`
- Article "Statistics and Social Network of YouTube Videos", Cheng, X., Dale, C. and Liu, J. (2008). IEEE International Workshop on Quality of Service, IWQoS. - `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.150.7896&rep=rep1&type=pdf`