Lastfm's Asian user network analysis

Pablo E. Duarte Tzuc
Data Engineering
Universidad Politecnica de Yucatan
Km.4.5 Carretera Merida – Tetiz
Tablaje Catastral 4448. Cp 97357
Ucu, Yucatan, Mexico
Email: st1809063@upy.edu.mx

***Abstract*—In the following document, we will work with the analysis of the social network Lastfm, a social network dedicated to the collection of data on artists or songs to find and give statistics to users. The network we are working with is only from Asian countries and contains a total of 7624 nodes and 27806 links, from this network we will see which are the important or interesting nodes in different aspects, for example which are the most important nodes in terms of number of connections, which are the nodes that can spread information more efficiently, which are the nodes that could be eliminated and lose connection of important information, etc. all this thanks to the centrality measures. Also, how the degree distribution can help us determine which is the most appropriate model for the network, whether or not there are hubs in the network and find the value of "degree exponent" (γ) and finally the communities that can be found in the network.**

***Index Terms*—Network, nodes, links, centrality measures, degree distribution, model for network, degree exponent, communities.**

I. Introduction

Last.fm is a social network in which users are responsible for providing data to the platform in order to obtain statistics and recommendations of new songs or genres, for example applications like Spotify or Deezer offer statistics or recommendations of new songs every season or a certain time, instead with last.fm we can enter data about our songs and music genres and get recommendations and statistics and even the data can be those of Spotify and Deezer just giving a few clicks on the application or website last.fm.

As for the data set of the graph was obtained from the Stanford Large Network Dataset Collection [1][2], this dataset was obtained by Benedek Rozemberczki and Rik Sarkar according to their paper "Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models" was obtained through an API. The network has 7624 nodes representing users from Asian countries such as Malaysia, Singapore, etc. as for the links it has a total of 27,806 which are the mutual relationships of followers among them. Possible applications include multinomial node classification, predicting the location of users and seeing if there are sets of users that are related.

Some of the characteristics of the network are that it is not bipartite because nodes cannot be divided into two disjoint sets V U and therefore no link connects a node U to a node V [3], From where the data was obtained, they mention that the network is not direct and that it does not have weights for the links, however, it was checked with a networkx function to corroborate this and effectively the network is not directed and the links don't have weights, it is not planar since the edges cross each other [4], it has an average degree of 7.2943 and a density of .0009568849118596328, so it means that the number of links is smaller than the maximum possible number of links, which means that pairs of nodes are not directly connected to each other [5].

II. Network Characteristics

- **Size of the network:** As I said previously the nodes represent the user of last.fm and we have in total 7, 624 nodes.
- **Number of links:** we reminder that the edges represent mutual follower relationships between users (nodes) and in this network there are 27,806 links.
- **Average path length:** The average path lenght of the network is 5.23, and it tell us how far a node is from other nodes [6]
- **Clustering coefficient:** and the clustering coefficient of the network is .219, it serve to to know how connected the neighboring nodes of a node are [6].
- **Diameter:** Perhaps we could think that from this graph that the longest shortest paths, have an extensive length, but it is not so, since the diameter of this graph is 15, that means that there are 15 links of distance between a node from one end to another, without forgetting that being a graph without weights, each link has the value of one [7].
- **Periphery:** I denote U = 1071, 3885, 2287, 2990, 4510 as the set of nodes that have the eccentricity equal to the diameter, it means that they are the nodes that have the longest distance between them and the other nodes of the network and at the same time have the same distance that the diameter of the network [8].

III. Centrality Measure

- **Degree centrality:** The ten nodes that have the highest degree centrality are denoted in the following A set, A = 7237, 3530, 4785, 524, 3450, 2510, 3597, 2854, 6101, 5127, This means that this set of nodes are the most

important in the network since they are the ones with the highest number of connections and since we are working with a social network this means that they have more influence, prestige or access to information [9].

- **Closeness Centrality:** the ten nodes with the highest closeness centrality are denote in the set E = 7199, 7237, 4356, 2854, 5454, 5127, 3544, 6101, 3450, 4900, so they are the nodes that spread the information very efficiently in the network [9].
- **Betweeneess centrality:** I denote the set B of 10 nodes with the highest betweenneess centrality, then B = '7199', '7237', '2854', '4356', '6101', '5454' '4338', '5127', '3450', '4785', this means that the nodes that are in set B are important, since if they are eliminated, an important part of the communication with the other nodes would be lost [9].
- • **Eigenvector centrality:** the ten nodes with the highest eigenvector centrality are in the set D = '7237', '3240', '3597', '763', '378', '2083', '1334', '3544', '4809', '2734', so these nodes have a far-reaching influence on the network [9].

I don't choose the page ranks because it isn't recommendable for undirected networks [9][10], also we can see that some nodes appear in different centrality measure, for example the node 3450 appear in degree centrality, closeness centrality and betweenness centrality, so it means that nodes is have many connection with other nodes, it spread the information quickly and if it nodes is eliminated disrupt the information with the other nodes, so in general it is a important node of the network. For other hand there are other nodes that do not appear like the node 3450, nevertheless it doesn't mean that the other nodes that appear in the different centrality measure aren't relevant.

IV. Degree distribution and models of network

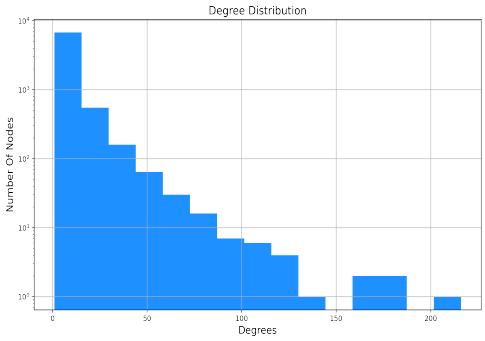

Fig. 1.

As we can see in the previous image, the distribution of the nodes are more skewed to the left, this means that most of the degrees of the nodes are relatively large as opposed to the degree range of the network [11], and this indicate that in the network are hubs, they are nodes that are highly connected to the other nodes of the network [11]. As we can see in the next imagen the node red is a hub.

V. Degree distribution and models of network

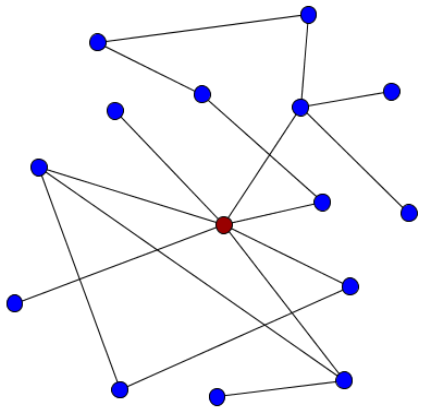

Fig. 2.

So to corroborate the before, we check the next plot of the network of Lastfm

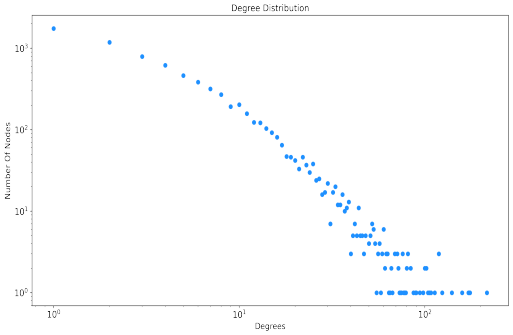

Fig. 3.

As we can see there a few nodes with highly degree, so these hubs are orders of magnitude larger in degree than most nodes and, this a characteristic of a scale free network [11][12]. Therefore, the most appropriate for this network is a scale-free network.

For to find the value of γ, we apply the next formula:

$$K_{max} = K_{min} N^{\frac{1}{\gamma - 1}}$$

Where:

$$K_{max} = 216$$
$$K_{min} = 1$$
$$N = 7624$$

Therefore:

$$216 = 1(7624)^{\frac{1}{y-1}}$$
$$\ln(216) = \frac{1}{\gamma - 1} \ln(7624)$$

$$\gamma - 1 = \frac{\ln(7624)}{\ln(216)}$$

$$\gamma = \frac{\ln(7624)}{\ln(216)} + 1$$

VI. Community detection

One of the main problems I had when finding the communities, was to use the right algorithm, in first instance I used the "girvan_newman" algorithm, but the main problem was the computational cost, because I left it running about 14 hours, in the morning of the next day I checked the algorithm was still running, so I discarded immediately to use this algorithm, then I used the algorithm "_naive_greedy_modularity_communities", but like the previous one this took too much time, so unfortunately I could not plot the communities provided by these two algorithms, in the end I chose to use the algorithm "greedy_modularity_communities", as this did not take long compared to the previous ones and could be rerun without problems, although this algorithm generated a total of 43 communities, so it was one of my main concerns, first because perhaps it was wrong and second because when plotting it would be a problem in terms of color, that is, there could be colors that are very similar or that could not be appreciated correctly, So I chose colors in such way that they won't seem.

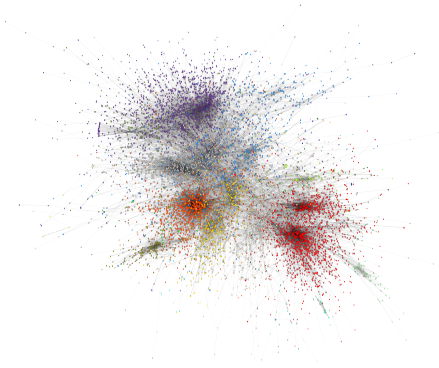

Fig. 4.

As we can see in the image above, there are only 6 of the 43 communities that can be immediately distinguished, that means that the other communities contain few users compared to the purple, low blue, red, orange, yellow and brown communities.

VII. Conclusion

Doing this work helped me to review some concepts in class, which are important, such as "degree centrality", "closeness centrality", etc. in general in the centrality measures, as these can give us interesting information about the nodes of the network, on the other hand the degree distribution can give us an idea of what is the most appropriate model for the network, as well as see if there are hubs in the network.

Without any doubt working with graphs that contain more and more nodes, could give us some inconveniences and visually it could be more complicated to obtain perspectives, for example this graph does not contain an abysmal number of nodes, however working with it and its graphs was not interactive, I mean that it was not easy to obtain conclusions in a visual way. Besides, I still have a lot to learn in this field, since there are many interesting things that can be found, but in my opinion, where this type of analysis can be used, or I would like to apply it more would be in the marketing sector or related to companies that sell products.

From my perspective it would have been great to see which Asian countries the communities belong to, to see if there are any interesting relationships, unfortunately the dataset did not provide information about the countries of the users.

References

[1] J. Leskovec, Stanfor university, . Available: http://snap.stanford.edu/data/.

[2] ROZEMBERCZKI, Benedek; SARKAR, Rik. Characteristic functions on graphs: Birds of a feather, from statistical descriptors to parametric models. En Proceedings of the 29th ACM International Conference on Information Knowledge Management. 2020. p. 1325-1334.

[3] ASRATIAN, Armen S.; DENLEY, Tristan MJ; HÄGGKVIST, Roland. Bipartite graphs and their applications. Cambridge university press, 1998.

[4] CHAPLICK, Steven; UECKERDT, Torsten. Planar graphs as VPG-graphs. En International Symposium on Graph Drawing. Springer, Berlin, Heidelberg, 2012. p. 174-186.

[5] Wikipedia, Wikipedia, 14 March 2021. [online]. Available: https://en.wikipedia.org/wiki/Sparse_network.

[6] STRANG, Alexander, et al. Generalized relationships between characteristic path length, efficiency, clustering coefficients, and density. Social Network Analysis and Mining, 2018, vol. 8, no 1, p. 1-6.

[7] WEISSTEIN, Eric W. Graph diameter. https://mathworld. wolfram. com/, 2003.P. R. d. l. Santos, Telefonica, 23 enero 2018. [En línea]. Available: https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/.

[8] CSERMELY, Peter, et al. Structure and dynamics of core/periphery networks. Journal of Complex Networks, 2013, vol. 1, no 2, p. 93-123.

[9] RODRIGUES, Francisco Aparecido. Network centrality: an introduction. En A mathematical modeling approach from nonlinear dynamics to complex systems. Springer, Cham, 2019. p. 177-196.

[10] A. shaw, strategic planet, 13 July 2019. [online]. Available: https://www.strategic-planet.com/2019/07/understanding-the-concepts-of-eigenvector-centrality-and-pagerank/.

[11] D. Q. Nykamp, Math insight, 23 05 2020. Available: https://mathinsight.org/scale_free_network.

[12] Barab AL and Albert R. Statistical mechanics of complex networks. Rev. Modern Physics, 7:47-97, 2002.