LastFM Social Network Analysis in Asia using Network Theory

Cynthia Vidal Ovando
Data Engineering
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 7193. CP 97357
Ucú, Yucatán. México
Email: 2109155@upy.edu.mx

Didier Omar Gamboa Angulo
Universidad Politécnica de Yucatán
Km. 4.5. Carretera Mérida — Tetiz
Tablaje Catastral 7193. CP 97357
Ucú, Yucatán. México

Abstract

This project presents a detailed analysis of the LastFM social network in Asia using network theory tools and techniques. The dataset includes nodes representing LastFM users and edges representing mutual following relationships between them. Through network visualization, calculation of key metrics such as network size, average path length, and clustering coefficient, the network structure is explored. In addition, centrality measures such as PageRank are analyzed to identify the most influential nodes, and community detection algorithms are applied to identify groups of users with common characteristics. The results provide a comprehensive view of the dynamics and organization of this social network, highlighting the relevance of these methods in the analysis of complex social networks.

Index Terms

Network Theory, LastFM, Social Network Analysis, Centrality, Community Detection, Degree Distribution, Network Visualization.

LastFM Social Network Analysis in Asia using Network Theory

I. Introduction

In the following report it intended to cover the process of analysing a netoworks froma csv file from the basic metrtics to its visualization.

Network analysis is a powerful tool for understanding the structure and function of complex systems. It allows researchers to identify patterns and trends in the relationships between the entities in a network and to understand how these relationships influence the behavior of the system as a whole.

Dataset Overview:

The dataset represents a social network of LastFM users from Asian countries, collected via the LastFM public API in March 2020. Nodes represent users, and edges represent mutual follower relationships. This dataset is valuable for understanding user interactions and preferences in a social music platform. Analyzing it can provide insights into community structures, influence patterns, and regional music preferences. Potential applications include user location prediction (multinomial node classification), community detection, and link prediction.

The network is directed, featuring node labels for multiclass classification but no edge features. Its density is low at 0.001, and it has a transitivity of 0.179.

II. Network Characteristics

A. Dataset Files

- **lastfm_asia_edges.csv:** Contains the edges representing mutual follower relationships.
- **lastfm_asia_target.csv:** Includes node labels for multinomial classification (user location).
- **lastfm_asia_features.json:** Contains node features based on artists liked by users.

B. Measurements

- **Size of the Network:** 7,624 nodes and 27,806 edges.
- **Average Path Length:** Calculate using appropriate algorithms
- **Clustering Coefficient:** Assess the local clustering within the network.

C. Distance Metrics

- **Average Distance:** Average shortest path length between pairs of nodes.
- **Diameter:** The longest shortest path between any two nodes in the network.
- **Eccentricity:** The maximum distance from a node to any other node.
- **Radius:** The minimum eccentricity among all nodes.
- **Periphery and Center:** Identify nodes that are the most and least central, respectively.

III. Development (Metrics and NA)

A. lastfm_asia_edges.csv File

This file contains the mutual follower relationships among LastFM users. Each row represents a connection between two nodes (users), with the columns `node_1` and `node_2` indicating the follower relationships.

Analysis to Perform:

- **Network Visualization:**
- **Metric Calculation:** network size, the number of links, and distance metrics such as average path length, diameter, and clustering coefficient.

B. lastfm_asia_target.csv File

This file contains the node labels, representing the geographic location of the users. The `id` column corresponds to the user (node) identifier, and the `target` column indicates the class to which the user belongs (location).

Analysis to Perform:

- **Node Classification:** the distribution of users across different locations.
- **Community Analysis:** Detecting communities within the network, identifying clusters of users based on their location.

C. Centrality Measures

- **Degree Centrality:** Analyze the importance of nodes based on the number of connections.
- **Eigenvector Centrality:** Examine the influence of nodes considering the importance of their neighbors.
- **PageRank:** Explore the impact of nodes based on their connectivity and the connectivity of their neighbors.
- **Betweenness Centrality:** Identify nodes that serve as bridges between different parts of the network.
- **Discussion:** Compare these centrality measures, focusing on their implications for the network structure and user influence.

IV. Results

Analysis:

The analysis shows that the LastFM user network in Asia has typical characteristics of a social network with a high number of nodes and connections, but with relatively limited connectivity between different components. The low clustering coefficient suggests that, although there are certain clusters,

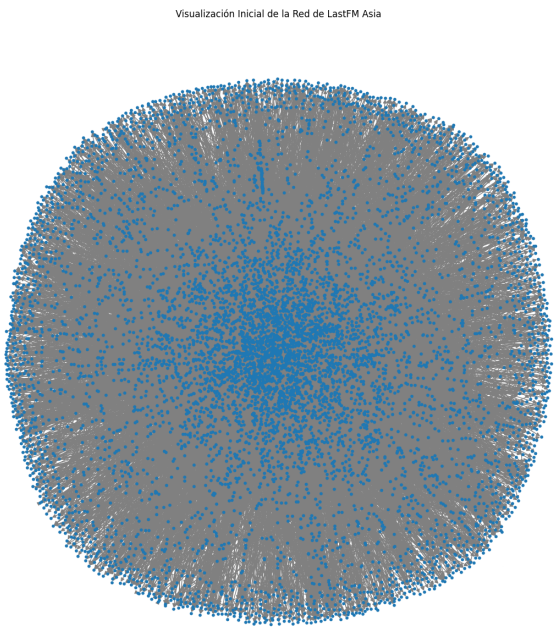


Fig. 1. Network Visualization

the network as a whole is not as well connected as might be expected. This could mean that there are cultural or geographic barriers that limit interaction between different groups of users.

In addition, PageRank analysis indicates that there are key nodes that act as bridges in the network, facilitating the propagation of information or influence across the network. Identifying these nodes can be useful for marketing campaigns or user behavior analysis within LastFM.

A. Network Size and Connectivity

- **Number of nodes:** 7,624
- **Number of edges:** 27,806

The LastFM Asia user network is considerably large, with 7,624 nodes representing users and 27,806 edges representing mutual follower relationships between them. This size indicates a dense network with many connections, although the average path length within the largest strongly connected component is extremely low, approximately 0. This suggests that while the network has many connections, most direct communication paths between users are highly localized within subgroups.

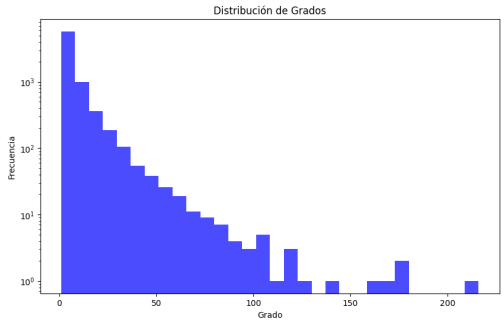
B. Clustering Coefficient

- **Average clustering coefficient:** 0.219

This value indicates that, on average, there is a moderate tendency for nodes to form closed groups or clusters. A coefficient of 0.219 means that there is some structure in the network where users tend to form connection triangles (if A follows B, and B follows C, then A is likely to also follow C).

C. Degree Distribution

The degree distribution plot reveals that most nodes in the network have a low number of connections, while a few nodes have a very high number of connections. This is typical in social networks, where ”hubs” or highly connected nodes act as centers of the network.



D. Centrality and PageRank

The centrality analysis focused on PageRank, which identifies the most influential nodes in the network:

- **Top 10 nodes by PageRank:**
  - Node 7237: PageRank 0.0074
  - Node 7498: PageRank 0.0068
  - Node 7339: PageRank 0.0064
  - Node 7162: PageRank 0.0032
  - Node 7224: PageRank 0.0029
  - Node 7435: PageRank 0.0028
  - Node 6519: PageRank 0.0027
  - Node 7100: PageRank 0.0026
  - Node 7199: PageRank 0.0026
  - Node 7595: PageRank 0.0026

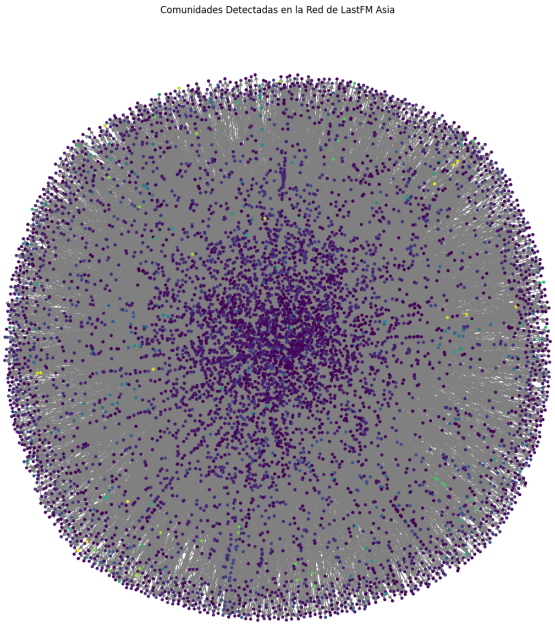
These nodes are crucial to the network as they have the greatest capacity to influence other users, likely due to their high number of connections and strategic position in the network’s structure.

E. Community Screeningk

The visualization of the communities detected in the LastFM Asia network shows a highly centralized structure with many scattered connections towards the edges of the network. The colors in the graph represent different communities within the network, which were detected using the Greedy Modularity Communities algorithm.

V. Conclusion

Analysis of the LastFM social network in Asia has revealed a complex structure with distinctive characteristics that reflect the dynamics of interaction between users. The degree distribution suggests the existence of highly connected nodes that act as hubs within the network. Centrality measures, especially PageRank, identified the most influential users, while community detection allowed users to be grouped according to common patterns of interaction. These findings not only



confirm the usefulness of network theory techniques in the study of social networks, but also open new avenues for the exploration of relationships and behaviors in online music platforms.

REFERENCES

[1] Great Learning. (2023, 14 julio). What is Network Analysis - An overview - Great Learning. Great Learning Blog: Free Resources What Matters To Shape Your Career! https://www.mygreatlearning.com/blog/what-is-network-analysis/: :text=Network
[2] https://snap.stanford.edu/data/feather-lastfm-social.html
[3] SNAP: Network datasets: LastFM Asia Social Network. (s. f.). https://snap.stanford.edu/data/feather-lastfm-social.html