Comparison of Network Analytics and Significance Analysis on Spotify Artist Feature Collaboration Network

Learning From Networks - Final report

Fabio Cociancich, Luca Fantin, Alessandro Lincetto

Master Degree in Computer Engineering - University of Padova

I. Dataset

For this project, we used the Spotify Artist Feature Collaboration Network from Kaggle [1]. This dataset consists of a graph where nodes correspond to artists and edges connect artists who have collaborated on at least one song. It has 156,422 nodes, which include around 20,000 artists who appear in the Spotify weekly charts and around 136,000 artists who had at least one feature with the chart artists, and 300,386 edges between them. Out of the information included with the nodes, the ones we used to analyze our results are the following:

- artist popularity, expressed as an integer number between 0 and 100 (100 corresponding to the most popular artist on the service), according to the Spotify API
- list of genres, according to the Spotify API

II. Code

All the code developed for this project, together with results files, dataset files and more, can be found in our GitHub repository [2]. The central script is `main.py`, which computes several graph- and node-level metrics on various graphs, depending on the command line arguments provided. The graph-level metrics include the number of connected components and a series of clustering coefficients (average, global, approximate global), while the node-level metrics include the local clustering coefficients and a wide array of centralities measures (degree, closeness, betweenness, PageRank, eigenvector). The script can work on the entire dataset, subgraphs taken from the dataset with respect to certain genres or popularity thresholds, and random graphs generated with the Holme-Kim model [3] with parameters specified through the command line arguments.

III. Data Analysis

In the first part of the analysis, artist rankings were calculated and sorted based on two key centrality metrics: degree centrality and closeness centrality. Degree centrality identified artists with the highest number of direct connections (i.e., those who collaborate with the most artists), so these artists occupy the top positions in the ranking. Some of the top-ranked artists include Johann Sebastian Bach, Traditional, Mc Gw, MC MN, and Jean Sibelius, who stand out for their high number of collaborations. Closeness centrality, on the other hand, highlighted artists who are more "central" in the network, meaning those who, despite having fewer direct collaborations, are well-positioned to interact with the entire network. Artists such as R3HAB, Snoop Dogg, Diplo, David Guetta, and Tiësto are at the top of this ranking, indicating their global influence within the network.

We have created a visual representation of the distribution of centrality measures for each artist. The resulting graph clearly shows how some measures, such as degree centrality, are widely distributed, with many nodes having a low number of collaborations and a few nodes having a very high number. Closeness centrality, on the other hand, tends to concentrate within a narrow range, suggesting that many artists are relatively close in the network, while some are significantly distant from the rest of the graph. Betweenness, PageRank, and eigenvector centrality show a similar distribution, with many nodes having low values and a few emerging with higher values, indicating that there are artists with very high influence in the network, but they are few. The comparison among the different centrality measures highlights how each of them measures different aspects of the structure and dynamics of the network.

Following this, the average ranking for each artist was calculated, based on the combination of rankings obtained through the various centrality metrics (degree, closeness, betweenness, PageRank, eigenvector centrality). This average ranking provided an overall view of the importance of each artist within the network, taking into account all dimensions of centrality. Artists like Snoop Dogg, Gucci Mane, and David Guetta emerged as the most important based on this average ranking, consistently ranking high across various metrics.

A complete ranking of the artists was finally calculated, based on all the centrality measures analyzed: degree, closeness, betweenness, PageRank, eigenvector centrality, and an average ranking that integrates the results of all the metrics. Betweenness centrality identifies artists who act as "bridges" between different areas of the network, such as Snoop Dogg, Gucci Mane, and David Guetta, while PageRank and eigenvector centrality highlight artists connected to important nodes, boosting their centrality value. The average ranking provides a comprehensive summary of the artists' importance, combining all the metrics. The analysis shows that some artists, such as Snoop Dogg and David Guetta, rank high in multiple rankings, suggesting their central and influential role in the collaboration network.

IV. Conclusion

A. Graph analysis

B. Genre subgraphs analysis

We analyzed also some subgraphs created considering only a particular genre. The genres with highest clustering coefficients are "latin" (0.165 avg. cc, 0.300 global cc.) and "trap" (0.189 avg. cc, 0.270 global cc.).

The ones with lowest clustering coefficients are "techno" (0.0017 avg. cc, 0.0029 global cc.) and "classical" (0.0013 avg. cc, 8.7 e-05 global cc., 1260 nodes, 775 edges, 541 connected components).

C. Popularity subgraphs analysis

By analysing the subgraphs created considering only the 0.1% most popular artists we can note that it has quite high clustering coefficients (0.277 avg. cc, 0.363 global cc.).

Contributions

References

[1] Julian Freyberg. *Spotify Artist Feature Collaboration Network*. URL: https://www.kaggle.com/datasets/jfreyberg/spotify-artist-feature-collaboration-network.

[2] Fabio Cociancich, Luca Fantin, and Alessandro Lincetto. *lfn_project*. URL: https://github.com/fantinluca/lfn_project/.

[3] Petter Holme and Beom Jun Kim. “Growing scale-free networks with tunable clustering”. In: *Physical Review E* 65.2 (Jan. 2002). ISSN: 1095-3787. DOI: 10.1103/physreve.65.026107. URL: http://dx.doi.org/10.1103/PhysRevE.65.026107.