CS176 - Project Proposal

Daniel Chukhlebov and Kyaw Soe Han
CS-176 Section 1

Spotify Artist Co-occurrence across Playlists in the US

Data Set found in this Repository (.graphML file)

The data set consists of a connected undirected graph with nodes representing an artist on Spotify and edges representing the co-occurrences of the two artists in public Spotify playlists. The weights on each edge is the sum of co-occurrence across the sampled playlists. The data only contains playlists created between Jan. 2010 and Nov. 2017 by users with a US-based Spotify account.

Out of the billions of playlists and millions of artists, the graph in the sampled data set contains 517 nodes and 18870 edges. This is for the sake of convenience since running calculations on millions of nodes on a personal computer is unfeasible. As a disclaimer, the sample size is significantly less than the optimal size to accurately represent the actual distribution of artist co-occurrences.

To ensure that we have a unique graph, we can run the setup code again and generate a fresh sample.

What we hope to learn from this analysis

Analysis of this dataset can provide us with a variety of useful insights on how artists and playlists relate to each other in a larger network, and form trends and larger structures.

For example, we can look into similarity between artists and genres. When people create playlists, they tend to group different artists together. We can examine the co-occurrences of artists across various playlists in order to find clusters of similar artists. We can also determine genre clusters, as artists that often appear in common playlists also likely produce music of the same or similar genre.

We can also plot out clustering and figure out how communities are tied together. To accomplish this, we will analyze how specific artists act as transitory edges between artist/genre clusters. Artists with high betweenness centrality likely form important connections between communities. This means that if users listen to artists from one genre, and also to an artist with high betweenness centrality, then they are likely to start listening to the other genre connected to the artist. This forms something similar to a Membership Closure, where a user following an artist related to a currently unknown genre is likely to begin interacting with that genre in the future.

In order to make various inferences about interactions between niche and popular artists, we will analyze factors like measures of centrality and how tightly clustered artists are. Using this information, we can come to conclusions about the tightness of groupings between popular/niche artists, the number of popular vs niche artists, and so on. We can also analyze how often popular artists appear in playlists alongside niche artists, and we can measure how strong the divisions between mainstream and niche music are.

To detect subcommunities within a homogenous group, we will use different methods like the Girvan-Newman algorithm. This can tell us about which artists in a broader community are most similar. In a business sense, this information could also lead us to decide which artists should be collaborating together on songs.

Works Cited

[1] Santos, Jonatas. “Spotify Network Analysis.” *GitHub*, 20 Jan. 2022, github.com/rodolfostark/spotify-network-analysis. Accessed 2 Oct. 2024.