CPSC 572 - Fundamentals of Network Analysis and Data Mining
Group 13 Project Proposal

Wasif Ud Dowlah - 30130706
Haoyang Shi - 30105296
Ty Irving - 30105319

Jan. 30th, 2023

The network of music collaborations: Exploring how collaborations influence music genres and artist popularity

Nodes and Links

Nodes: Each node in our network graph will represent an individual music artist.
Links: A link is formed when two or more artists have collaborated on a song or an album.

Dataset

Our project is centered on mapping the collaborative networks among music artists across a variety of genres. To compile this dataset, our strategy includes acquiring data directly from sources such as Kaggle.com, complemented by the use of web scraping techniques to gather additional information from platforms like Spotify and other music-related databases.We anticipate that the data collection process will take approximately two weeks. A primary challenge we anticipate is the need for extensive data cleaning and preprocessing to ensure uniformity and compatibility between the different data sources. For web scraping (if necessary), data cleaning, and handling API requests, we will use Python. We expect the data to be initially stored in a tabular format, either CSV or JSON, and it may require additional processing to convert it into a network format suitable for our analysis. We will implement machine learning algorithms for pattern recognition and network analysis. This includes using Python libraries like Pandas for data manipulation, and NetworkX for network analysis.

Expected size of network

The expected size of our music collaboration network is approximately 500 to 1000 nodes, with a corresponding number of links. This range is chosen to provide a suitable analysis while maintaining computational efficiency. In order to build this network and have a reasonable overlap of artist collaboration we will be choosing a certain subset of artist in order to create a network that will yield us useful results since if we take a broad range of artist from all over there is likely not going to be that much overlap within our network making it very unconnected which is why we have to take a subset of artists.

Questions you plan to ask and why do we care?

Our network is significant because it can inform industry stakeholders about emerging trends, guide new artists in strategic collaborations, and provide valuable insights into the dynamics of the music industry. Here's what we are curious about:

1. Which music genres are most influential in terms of collaborations?

2. Which artists are the most influential in terms of collaborations?

3. Does increased collaboration correspond to an increase in popularity?

To address these questions, we plan to use centrality analysis, specifically degree centrality to find influential artists in the network. Centrality measures will help us identify the most important nodes in the network based on their connections and the connections of their neighbors.