Playlist creation with GNNs

Authors: Alič, G. (watermelon), Drusany, J. (blueberry), Piciga, A. (strawberry), Ploj, Y. (banana)

# Introduction

Music streaming services have become increasingly popular over the past decade, with millions of users streaming music every day. However, one challenge faced by streaming services is creating dynamic playlists that are tailored to the users' preferences. In this project, we will explore how graph neural networks (GNNs) can be used to generate playlists that are tailored to a user's preferences.

# Dataset

For this project, we will be using Spotify's Million Song Dataset, an open-source dataset containing information about over one million playlists. The MSD contains various metadata about the songs, such as artist name, album title, and song duration. The dataset contains around 1 million playlists, 2 million songs and 300 thousand artists. Since the dataset is way too large for us to process entirely, we are going to sample a subgraph by taking only the most popular songs and playlists (according to the highest node degree).

# Graph structure

We will use this dataset to construct a heterogenous graph, with nodes representing songs, artists and playlists, and edges representing authorships/playlist inclusions. We will then use a GNN to generate playlists for each user based on their past recorded preferences.

# Graph ML model

We will use GraphSAGE, a graph neural network architecture for semi-supervised node classification, for our GNN model. GraphSAGE is a neighbourhood-based graph neural network, where each node’s representation is aggregated from its neighbours. This model is appropriate for our task because it can capture the similarities between songs and generate playlists that are tailored to a user’s preferences.

Input: a playlist with some songs.

Output: the playlist with additional songs.

Graphical user interface, text, application

Description automatically generated

# ***Image source:***

# *Some examples of meta functions for the GraphSAGE model:*

# *$$h\_i = [1]$$*

# *$$h\_i = [\text{duration}(i)]$$*

# *$$h\_i = [\text{md5}(\text{title}(i))]$$*

# *or some other string to vector representation.*

# *$$h\_i^{(l+1)} = \text{ReLU}(\sum\_{v\in N(i)}h\_u^{(l)})$$*

# *$$h\_i^{(l+1)} = \arctan(\frac{1}{|N(i)|}\sum\_{v\in N(i)}h\_u^{(l)})$$*

# Evaluation metric

We will train our model by hiding edges and trying to predict them back. Our evaluation metric will be the accuracy of these self-supervised training predictions.

# Conclusion

In this project, we will explore how GNNs can be used to generate playlists tailored to a user’s preferences. We will use the Million Song Dataset to construct our graph and GraphSAGE as our GNN model. We hope that our project will provide insights into how GNNs can be used to generate personalized playlists.