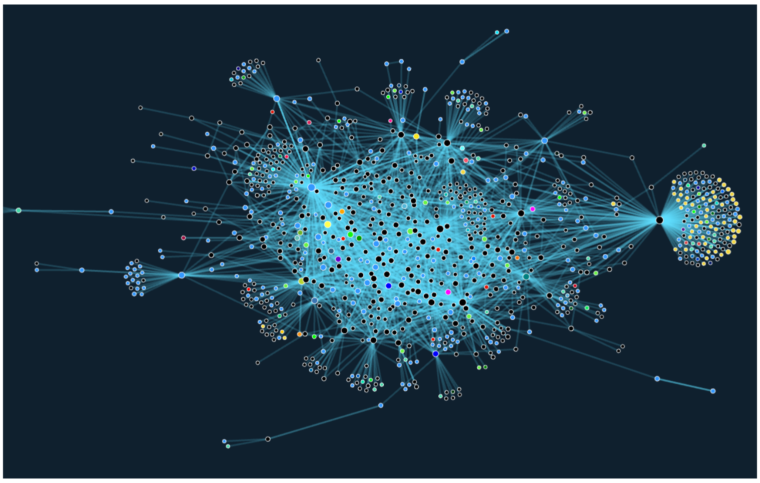
École Polytechnique Fédérale de Lausanne

A Network Tour of Data Science

Movielens 100k Dataset Project



Team 15
Arthur BABEY
Stanislas DUCOTTERD
Nessreddine LOUDIY
Lamyae OMARI ALAOUI

Lausanne - Switzerland

Fall semester- 2019

1 Purpose

The main purpose of this project is to build a recommender system: given a user, can you predict the rating that it will give to a given movie? In order to give a prediction we will use graph neural networks to do matrix completion.

2 Dataset

Movielens is a personalized movie recommendation system. Several datasets have been built using this database, the smallest being Movielens 100k. It contains 100,000 ratings from 1000 users on 1700 movies which means that almost 95% of the values are missing in the matrix. Various kinds of information are available about the users (age, gender, occupation, zip code) and the movies (release date, genre).

The movie features can be easily expanded if needed as there is a lot of available APIs allowing us to get the list of main actors, the budget and other features to help build the network.

3 Network

We have two networks; the users network and the movies network. For the first one the users are the nodes and the edges will be built from the features (age, gender, occupation, zip code). In the second graph the movies will be the nodes, the edges will also be built from the features (release date, genre).

4 Goals

The main goal is obviously to build the recommender system that will complete the missing ratings in our dataset. To achieve this, we will need to:

- Build a meaningful graph for the movies and for the users
- Train a neural network to complete the matrix and use the two graphs to constrain the space of solution.

Besides our main goal, we can try to analyse how the ratings are distributed over the US states using the zip code, and how the age, gender and occupation affect the users rating for the different movie categories.