Name: Marek Furka **Student id:** 11934086
Main supervisors: Marcia R. Ferreira, Niklas Reisz
Co–supervisor: Prof. Allan Hanbury

Interdisciplinary project in data science - project proposal

Introduction:

Many datasets these days can be analyzed in a form of a network where some entities (nodes) are connected by some relationships (edges). In this project, the domain to be explored is music. There are some phenomena that can be observed in many domains such as preferential attachment (new entities are more likely to be attached to the more popular ones). Another common property is forgetting (older entities become less popular over time), and "evergreen" phenomenon (some entities are not forgotten over time in contrast with the forgetting property). The idea of the project is to analyze how these properties and characteristics of the "evergreens" differ from other well researched domains such as scientific papers citation networks [1].

Goal:

The goal of the project is to explore, preprocess the dataset, to perform general network analysis and most importantly to analyze whether the phenomena commonly observed in other domains that were explained before can also be observed in this domain. It is expected that the characteristics of the "evergreens" are very domain specific and will differ from other domains. Additionally, the aim is to analyze whether the identified characteristics differ significantly among subsets of data when taking into account different age groups or countries of users in case the data quality and size will be sufficient after preprocessing and exploration. Similarly, if there will be enough data available, the plan is also to try joining other suitable datasets to the main one, such as a dataset containing information about song's genres, so that we could investigate if the characteristics being analyzed significantly differ for different genres of music.

Data:

The data originates from last.fm which is an online service that tracks people's music listening habits. The particular dataset is called "lastfm-dataset-1K" [2] and contains information about all the song listening's of a subset of nearly 1000 users during the years 2005-2009. The attributes consist of user id, timestamp, artist id/name, track id/name. Additionally, some data about the users is provided in a separate file – it contains information about gender, age, country and date of registration for most of the users.

Approach:

The project will consist of the following tasks with their effort estimates indicated in the brackets:

- Data set preparation (5 hours)
- general data exploration, check for anomalies and other possible problems in the data (10 hours);
- calculate and analyze network metrics (10 hours);
- analysis of the preferential attachment phenomenon (25 hours);
- analysis of the forgetting phenomenon and characterization of the "evergreens" (35 hours);
- analysis of "evergreens" and forgetting phenomena for different age groups/ countries of users (15 hours);
- exploration of adding information from other datasets and analysis of the effects of the additional attributes on the "evergreens" and forgetting phenomena (e.g. song genre/ artist's nationality...) (15 hours);
- writing of report (10 hours).

Expected outcome:

The expected outcome is an extensive analysis of the network properties/phenomena commonly observed in other domains. Part of the project's output will be basic data preprocessing and loading which can be together with the analysis be used as a baseline for future work such as simulation and prediction of the network's behavior for example.

References:

[1] Newman, M. 2001. Clustering and preferential attachment in growing networks. Physical Review E, 64(2)
[2] http://ocelma.net/MusicRecommendationDataset/lastfm-1K.html [Accessed March 13 2021]