# Network Tour of Data Science: Project Proposal Course Suggester

Cherqui Alexandre, El Hamamsy Laila, Gusset Frédérick, Triquet Thomas

## 1. Introduction

Once the master level is reached, students have the opportunity to choose between a vast pool of courses, as well as courses which are not part of their original study plan. However, aside from going through all the coursebooks or asking other people's opinions, there is nothing to help students select courses which may be relevant to their domains of interest. That is why we had the idea to work on a course suggester. A student would be able to input the name of a course of interest or a list of courses taken and obtain a list of courses which are related.

## 2. Data Acquisition, Exploration and Cleaning

To do so we would need to create a graph where the nodes are the courses and the links are given by features recovered from multiple EPFL databases. The idea is to make use of the list of courses that students took as well as the course descriptions to weight the edges. From the latter we would be able to recover multiple attributes : the professor, keywords, pre-requisites, teaching methods, assessments, the faculty to which the course is attached and so on. All of this information is public either via is-academia or the public access.

The courses each student has taken since 2001, their affiliation (MT, EL, IN...), the current semester (BA1, BA2,...) as well as an XML with the course descriptions were provided to us by Francisco Pinto who already distributed this dataset during the DataJam days. This can be found on our github here. It is important to note that the data is anonymous so there will be no way to trace back any information to a particular student. Therefore we shouldn't have to do any web-scraping but there will be a large portion of data cleaning and formatting before being able to construct the graph. For example, we may need to match the courses over the different years because their names might have changed and handle the potential change of language along the years. We have already started cleaning the data for both the enrollments and course descriptions as can be seen on *our github*.

## 3. Feature Selection, Graph Construction and Analysis

To construct the graph, we will have to determine which features are the most relevant for the desired application. Then, the goal would be to find clusters of courses that are related thanks to the previously described features. Thus, one could find all courses which would have strong connections to a given course. A person could then input a course and would receive a list of relevant courses which they could take. Another possibility is to give a list of courses of interest or courses previously taken.

To extract the list of relevant courses several ideas came to mind. One idea is to create a subgraph would from the inputs, and the most probable features to describe this cluster would then be isolated. Starting from this new dynamic feature selection, a new graph could be constructed and new clusters could be extracted from it in order to output the best courses for the given input. This is close to the principal of template matching. Another option is to use what was in the last slides of the Graph Signal processing course with the recommender systems.