NTDS Project - Team 1 : Project Sumarry

Authors : Magnin Jonathan, Nonaca Darja, Shmeis Zeinab, Wang Shu Date : December 4, 2019

The goal of our project is to investigate the gender and nationality distribution over EPFL campus (i.e. in each section). Observing this distribution can reveal whether there exists a correlation between the students nationality/gender and their choice preferences. We use network science because it is able to create the synergies we want between sections and which can not be obtained by simple statistics methods that strictly separate sections.

For the dataset, we use our own dataset parsed from *www.epfl.ch/campus/services/ressources/is-academia/acces/accesspublic-bachelor-master/* (into CSV files. Note that the EPFL login is required for complete data). This gives us access to semestrial lists from 2012 to 2019 containing : names, SCIPER, gender, nationality, section, courses, professors of the courses, section of the courses, academic semester.

We chose to explore these data by building graphs of students and courses :

	Students graph	**Courses graph**
Nodes	Students	Courses
Features	Courses	Students
Signal	Nationaility / gender labels (discrete)	female ratio (continuous)

We will build such graphs for every year from 2012 to 2019. Visualizing them will give the information we are searching for and could also reveal changes from a year to another.

We will exploit these data in the following way :

- Create the epsilon-similarity graphs through RBF kernel
- Use dimensionality reduction and clustering to create visualizable graphs that are consistent with EPFL structure (section clusters). We will label graphs with sections in order to tune the graphs and reduction parameters until the final graphs are consistent with EPFL structure (sections).
- Label graphs with nationality, gender and female ratio to vizualize the distribution, use graph filtering to remove outliers and build a general picture.
- Use Gephi in order to obtain good and easily readable final graph visualization.
- Use machine learning to predict gender and nationality of a given student knowing its courses.

Finally, we will synthetize our observations to give a summary of main nationality preferences and gender distribution. We will also try to put in evidence a "female gradient" over EPFL sections.