Project proposal

Claudio Loureiro
Cem Musluoglu
Jordan Willemin
Cyril van Schreven

1. Chosen dataset

For this project, we will use the dataset from Kaggle: “Stackoverflow developer survey, 2017”[1]. Stackoverflow is a forum focused around programming. The dataset contains information about 64 000 developers active on the forum, who accepted to fill in the survey. The survey has more than 100 questions, all of them are multiple choice. The questions range from country of residence, formal education and known coding languages, to ‘tabs or spaces?’ and ‘is it OK to use loud clicky keys in an office?’.

In our network, nodes will be the users and links will be the similarity between two users.

2. Possible outcomes

We will start by making graphs, visualizing the data, looking at clusters and then see what information or correlations we can extract from this survey.

In the networks we intend to find which questions distinguish the users of the community.

Do coding languages influence the salary, but also the habits and the ethics of the participant?

Which combination of questions has the most information about the nationality of the programmer?

[1] https://www.kaggle.com/stackoverflow/so-survey-2017/data