TASK 3 – Social Network Analysis (15%)

In the computer networks research group, several of our researchers deal with networking problems in online social networks (OSNs, e.g., Facebook or Twitter). Research in these networks is a relatively young area that has gained a lot of popularity in the past five years. Problems range from building systems for these networks to analytical tasks like predicting link establishment between OSN users.

In this context, one recurring concept is that of influential users in the network. These users are important in many applications and research proposals as they have an impact on how links are formed, content is distributed, opinions are built, advertisements are perceived and so on.

Therefore, various stakeholders are interested in identifying these users. In this third task, you will try to help them to some extent: Based on some real-world Twitter data, you are requested to build a model that can decide, if given a feature vector of two users (user 1 and user 2), which of these users has more influence on the network.

Download the dataset here.

You will find two different sets, a training set, and a test set. To obtain comparable results among all participants, please make sure that you use the training set to develop your algorithm (see instructions below) and use the test set to evaluate your algorithm.

Note the following: in the dataset there are three features for each user that describe the user's network structure (e.g., how her followers are connected). The exact meanings of these features have not been given to us by the data owner, but you can interpret them as something closely related to metrics like betweenness centrality. Bonus points will be given if you can make a convincing case what the network features could represent (note: optional and not mandatory to earn the highest grade). All other features should be self-explanatory.

In this task, there are no hints. Your task here is to:

1. **Analyze the dataset to obtain helpful information about the data.**
2. **Fill in missing data with a method of your choice (do not simply delete these samples).**
3. **Based on your analysis, you should build a ML model that yields a high accuracy to determine which of two input users has more influence on the network.**
4. **Present your findings in class.**
5. **At the end of the semester, summarize your findings together with those of the other tasks in the final report.**