# Project Summary: Spammers on Social Networks

## Students

Team 04:

- Yann Yanis Bouquet
- Ali El Abridi
- Tariq Kalim
- Jonas Müller

## Motivation

The goal of the project is to implement a model/algorithm capable of classifying users of a social media network as "spammers" or "not spammers" based on their interactions with other users and a limited set of features.

## Dataset

We opted for the "Social Spammers" dataset from UC Stanta Cruz. The dataset is a graph of interaction between the users of the social network Tagged.com. The data is organized as follows: We have a dataset containing 5.6 million users (23.2 MB: userId, sex, timePassedValidation, ageGroup, Spammer/Not-spammer label) and a dataset containing 856 million interactions between users (6.8 GB: day, time_ms, src, dst, relation). These relationships correspond to links between users. These relationships have 8 different types filled in by a "relationship" ID.

## Graph

Unsurprisingly, the vertices are the users of the website and the edges are the interactions between them.

## Challenges

- Size of the dataset: due to its large volume (>6 Gb), the dataset can't be processed easily as is. Sampling methods will be necessary.
- Features: Apart from the types of interactions, there are only 4 features tied to the users, meaning "classic" classifying approaches might not be good enough

## Engineering

- Downsampling: This essentially means that we have to find a representative subset of the vertices. However, since the network is connected this will mean to break/neglect edges which in turn introduces an error.
- Extract graph features: Due to the structure of the dataset more features can be extracted from the different interaction graphs. To begin with the deegres, the clustering coefficients and the number of triangles can be added as features to explore their correlation with the "spammer" label.
- Classification: To classify the constructed feature graph we then have to deploy a classification algorithm that works best with the given data structure. As to start with we might compare the linear vs. nonlinear methods learned in class with each other (PCA, MDS vs. Isomap, Laplacian Eigenmaps, LLE).

## Proposal

In order to perform the data acquisition, we must select a subset of the data to reduce the excessive size of the initial datasets.
The data organization invites to build several graphs according to the number of types of relationships that we will have kept after reducing the size of the dataset. If we create a directed-graph per relationship, we will have each graph representing users as nodes, and the relationships between users as links. In this way we will be able to generate features with different graph analysis methods: Pagerank, Degree, k-core, Connected Components, Triangle Count... In this way we will be able to use the data to compare the graphs according to the type of the associated relationship. These comparisons can be made at the graph properties level, their identified types, node properties and attribute analysis. We will also visualize the graphs in order to obtain a qualitative analysis. These analyses will allow us to develop operating methods to compare different classifiers using the tools seen in progress according to the chosen relationships and graph analysis methods. These methods will initially be based on existing work on the detection of spammers: SVM, K-NN.