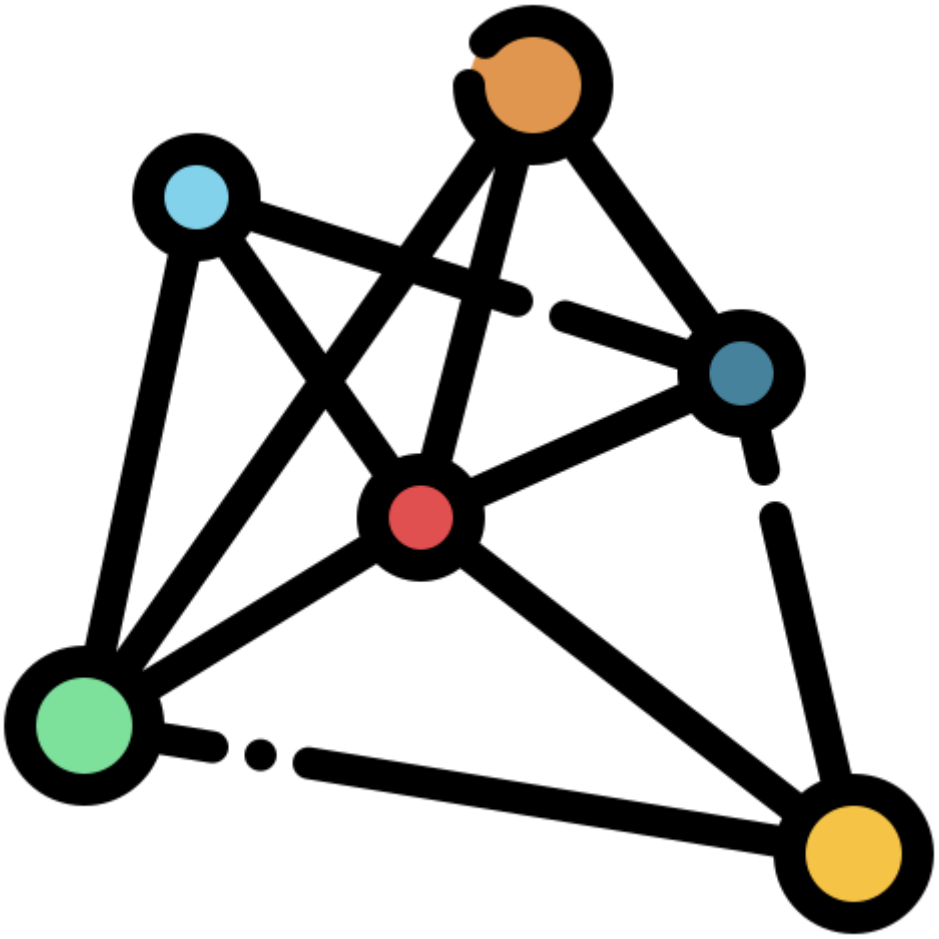
Challenge 1 :
The illusion of hidden personal data



Teachers :
BOTHOREL Cécile cecile.bothorel@imt-atlantique.fr
HARALAMBOUS Yannis yannis.haralambous@imt-atlantique.fr
JULLIEN Nicolas nicolas.jullien@imt-atlantique.fr

Students :
LE MÉE Aurel aurel.le-mee@imt-atlantique.net
SAILLY Gregory gregory.sailly@imt-atlantique.net

Table of Contents

Introduction

LinkedIn aims to conduct a marketing campaign for a restaurant in the San Francisco Bay Area to promote its business. To do this, the team decides to identify the 5 most influential users of the social network who could help disseminate information about this marketing campaign. However, only 40% of users have provided their location in their profile, making it difficult to identify relevant influencers for this campaign.

Business Problem

Definition of the Graph

The graph available to the marketing team consists of the set of users and their relationships.

- The nodes represent the users
- The edges represent the non-reciprocal connection of one user to another.

By definition of the edges, it is understood that this is a directed graph without weight. That is, a user can follow another without the latter following the first.

Definition of influence

A user's influence is defined as their ability to influence other users of the network. It can be measured using many metrics such as the number of followers, the number of content shares, the number of mentions, etc. In the context of a marketing campaign for a restaurant in the Bay Area, a user's influence can primarily be measured through the following characteristics:

- The number of connections:

The more connections a user has, the more likely they are to have an influence on a large number of people.

- Location :

For a marketing campaign, the influencer's location is a very important parameter. Indeed, an influencer located in Europe will probably have very little influence on a restaurant located in San Francisco. However, more than the influencer's location, that of their connections may be even more important. Indeed, if the influencer is located in Europe but has a majority of connections from San Francisco, the latter will probably have more influence than if he was located in San Francisco but with connections whose origin is mostly European.

- Centrality :

The user's centrality in the network is also a very important element. Indeed, if the user has a large number of connections but these in turn do not have a minimum number of connections, then the user's influence will certainly be limited. Whereas if the latter is at the center of the network, his influence can more easily spread throughout the network.

- Homophily :

This parameter can be very important both to retrieve missing information from some users and to influence their connections. Indeed, when homophily is high, it means that users have very similar characteristics. Consequently, they probably know each other very well and the proximity of users can be a good factor of influence since a user will more easily trust a person they know rather than the opposite.

- Prestige

The prestige of a user's background or career can have influence. Indeed, despite not being related to the world of restaurants, a user whose academic background and career are prestigious will undoubtedly tend to be followed more easily. Moreover, if their career is related to the world of restaurants, this is all the more important.

Definition of marketing campaign

The marketing campaign consists of promoting the restaurant to LinkedIn users in the San Francisco Bay Area. The campaign will be considered successful if it reaches a certain number of shares, views on the social network but also if it has had a certain influence on the restaurant's revenue.

Quantification of success

The quantification of success will be done in two stages.

- The first will correspond to the accuracy with which the missing data was predicted.
- The second, for its part, will be interested in the influence that the 5 influencers chosen for the marketing campaign have on their network and on the restaurant's revenue.

Not having access to the campaign results. The quantification of the success of this stage will be done at the time of determining the 5 influencers, notably using the influence parameters mentioned previously such as their number of connections, their location, that of their connections and the centrality of the influencer in the network for example.

Optimization problem

In conclusion, the problem is to identify the 5 most influential LinkedIn users in the San Francisco Bay Area in order to promote the restaurant's marketing campaign. To solve this problem, we will use a profile completion strategy including homophily to complete incomplete profiles. We will then use network analysis methods to identify the most influential users based on the metrics defined above. Finally, these results will be evaluated using metrics such as precision, recall, F1 score, for example, in order to measure the relevance of the predictions and to be able to decide if our solution meets expectations.