

|  |
| --- |
| Auteur1  [auteur1@](mailto:auteur1@i)imt-atlantique.fr  Auteur2  [auteur2@i](mailto:auteur1@telecom-bretagne.eu)mt-atlantique.fr  Auteur3  [auteur3@i](mailto:auteur1@telecom-bretagne.eu)mt-atlantique.fr |

|  |
| --- |
| Titre du document Rapport  Version 0.1 - 01/01/2017  Formation d'ingénieur  Année scolaire 2017-2018 |

Sommaire

[Résumé 3](#__RefHeading__20344_1162629425)

[Abstract 3](#__RefHeading__20346_1162629425)

[I. Introduction 6](#__RefHeading__20348_1162629425)

[II. Le projet 6](#__RefHeading__20350_1162629425)

[II.1 Aspects techniques 6](#__RefHeading___Toc1855_245993825)

[II.1.1 Le logiciel 6](#__RefHeading___Toc1857_245993825)

[II.1.2 Le hardware 6](#__RefHeading___Toc1859_245993825)

[II.2 Aspects économiques 6](#__RefHeading__20462_1162629425)

[II.2.1 Coûts de production 6](#__RefHeading__20572_1162629425)

[II.2.2 Etude de marché 6](#__RefHeading___Toc1861_245993825)

[III. Conclusion 6](#__RefHeading__20352_1162629425)

[Références Bibliographiques 7](#__RefHeading__20354_1162629425)

[Annexes 9](#__RefHeading__23102_1162629425)

[Annexe 1 10](#__RefHeading__23104_1162629425)

[Annexe 2 10](#__RefHeading__20574_1162629425)

# Introduction

As data scientist at LinkedIn, our colleagues from Marketing want to organize an online marketing campaign for their client, a restaurant in the Bay area. They have read a lot about social network analysis and asked us to find the 5 most influential people on the network who would best promote the restaurant.

When taking a closer look at the dataset we are given, we can see that only 40% of users have provided their location. We will therefore have to fill the uncomplete profiles before finding the influencers.

In order to do that we are given a dataset representing a LinkedIn graph, each node being a person and edges the fact they are connected. People also have three attributes: location (locations they live/have lived in), college (college(s) they went to), employer (company(ies) they worked for). Only 40% of the people in the graph have provided their attribute.

Our aim is to use connections and correlations between those attributes and people in order to fill the missing profiles, and use those results in order to find influencers.

# problem enunciation and statistics.

## problem enunciation and task analysis



As data-scientists, we intend to use the CRISP-DM methodology in order to solve the problem, which consists in different phases as presented in the chart beside.

First of all, we needed to do some researches about similar studies, because it may help us understand more the problem and see different solutions. We used several papers but the most useful was [1] User Profiling in an Ego Network: Co-profiling Attributes and Relationships.

We should now list all the factors we think might lead a user to have a certain attribute, as we will use it later. This will help us understand the statistics that are relevant to draw from the graph.

We then have to understand how the dataset is made so that we can use the data properly.

We then draw the statistics and use the statistics to elaborate one or several models for user profiling.

Once it is done, we have to implement the model using the data given and assess the model by comparing our predictions with the ground truth. Once it is done for every model, we will know which one are the best, meaning the one with best accuracies.

We will now have the filled profiles and will be able to use it to establish who are the influencers that might help us find a restaurant at the bay.

## Statistics

We therefore want to draw interesting statistics from the graph that might help us understand what factors make that a node has a certain attribute.

First some basic statistics about the graph:

*Number of users in our graph: 811*

*Number of users with one or more attribute college: 230*

*Number of users with one or more attribute location: 336*

*Number of users with one or more attribute employer: 297*

*We have to find information for 475 users with empty profile*

*Number of edges: 1597*

*Average degree: 3.9383*

*Graph density 0.004862*

*Diameter (maximum eccentricity): 19*

*Radius (minimum eccentricity): 10*

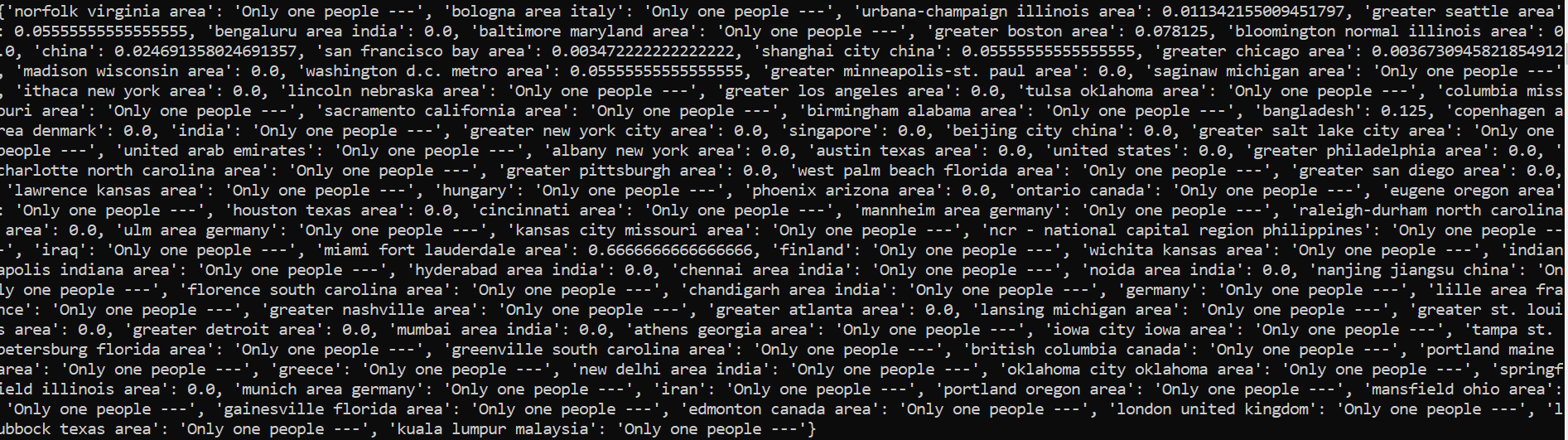
*Mean eccentricity (eccentricity(v) = the maximum distance from v to all other nodes): 13.241676942046857*

*Center is composed of 3 nodes (nodes with eccentricity equal to radius)*

*Periphery is composed of 5 nodes (nodes with eccentricity equal to the diameter)*

*Mean clustering coefficient 0.347288*

*Total number of triangles in graph: 1217*

* We might also want to know, for each location for example, the clustering coefficient for the graph composed only of the nodes that have this value for attribute location. For example, we look at all the people living in location L1 and compute the clustering coefficient for the location L1. If values are high, then we may be able to determine easily whether a user lives in this location or not. Here are the results :

We can see many locations have only one person in it and if they have more than one person, the clustering coefficient is really low. We therefore cannot use this information in our prediction model.

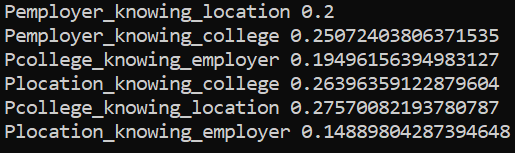
* We can compute the homophily as well:

The homophily for Employer, Location, College are respectively: 0.0338, 0.04320, 0.03569.

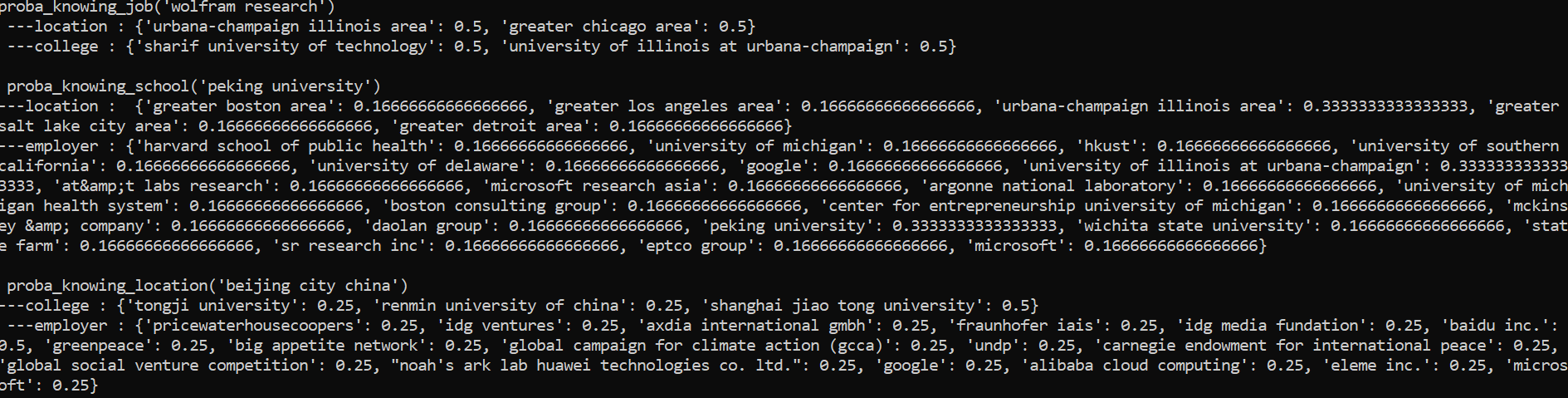
Theses values are low due to the fact very few people have completed their profiles!

* We also wanted to know if sharing an attribute might help sharing another one. For example : if i and j are in the same college, what is the probability they have the same employer ?

Here’s our result



We can see the ¼ of the people from the same location share the same college and vice-versa, which is an interesting point for the models we might want to implement.

Other interesting conditional probabilities would be the probability for users with the same attribute (ex : location) to have another attribute ( ex: employer ). Using it, we might be able to know, for example, that every people who have been to a certain college have a certain attribute location, etc. Below is a sample form our results (for a given job, given college and given location)

We can see this information can be useful, because sometimes, knowing an attribute of a user might help us define the most probable values for his other attributes.

# PREDICTION MODELS AND HYPOTHESeS

## PREAMBULE : ACCURACY EVALUTATION

We will need a precise way to compute the accuracy of our model, by comparing the predictions with the ground truth. Our aim is to fill the profile, but for an attribute, there can be several values (mostly there is only one value but it isn’t impossible that there are more than one value). In a first time, we predict only one value for each attribute, but later we might predict several values.

The evaluation process will be done this way : we will compare the predicted results and ground truth, and return the number of correct predictions divided by the number of predictions.

## THE NAIVE METHOD

We can use the notion of homophily to help us begin solving this problem. Indeed homophily is the ‘tendency of individuals to associate and bond with similar others’ (Wikipedia). If similarities breed connections, we might assume that a proportion of people that are connected are connected because they have similarities. From a logical point of view:

We have ‘A & B sharing attributes => A probably linked to B’ (with a probability equals to the homophily for the attributes shared)

Which is equivalent to ‘A,B not linked => A&B are unlikely to share attributes’.

What we want to do is consider that people who are connected are likely to share one or more attribute, which is not directly equivalent to the concept of homophily but a hypothesis.

Considering this, the first idea that came through our mind was, for a given node, to predict the value of one of his attributes by the most represented value amongst his neighbours’ attributes. (ex: I have 5 neighbours, 3 of them went to college C1, 2 of them went to college C2, our model will predict that I went to college C1).

Results :

The results are not very good. It is logical, if we consider the statistics (for the medium-size network) : the average degree is ~4, which means the sample we use for the prediction isn’t big enough. We need to think how we can extend it without decreasing the precision (which might happen if we use a too large sample) .

## COMMUNITY DETECTION BASED METHOD

After seeing the results from the naive method, we’ve thought about extending the method to a larger part of the network. The next idea that came to our minds was that we could detect communities from the graph. Once we have our partition of the graph and the communities, we know in which community the user is and can predict the value of one of his attributes by the most represented value amongst the people in his community’s attribute.

In order to make the partition and split the graph into different communities, we used the Louvain algorithm and the heuristic provided by the package NetworkX:

*partition = community.best\_partition(G)*

We then had to look what community each unfilled node was in, and browse through the community, take the number of times each location is represented in the community and predict the location corresponding to the most represented area. We can draw a list of the occurrences for each location, compute its maximum and standard deviation, and consider as probable all the locations with occurrence superior or equal to ‘*maximum – standard deviation’*.

The accuracy is : 

The accuracy is the best for college, and the least for employer.

In order to increase the accuracy, we tried to do the prediction by browsing through both the community and the first level ego network. By doing it we’ve had



We can see the prediction is much better for college and location but is worse for the employer.

We will therefore use it only for the prediction of the college and the location.

## USING CONDITIONNAL PROBABILITIES

After drawing the statistics, and a few experiments on the data we’ve had the idea to use one’s known – or predicted – attribute to predict his other attributes.

As college is our most accurate prediction, we wanted to use the predicted or known value of a user’s college in order to predict his location and employer.

For a given user, we take his value for attribute college. If it is missing, we take the predicted value(s). We search for everyone that has been to this college (let’s call it C1) and store their locations. For each location in the stored locations, we compute the probability of living there knowing we went to C1. We then have a list of all the locations where the people who also went to this college usually live in and the probability (Pli they live in this location.

In order to do the prediction, we used the community detection- based method and took all the occurrences multiplied them by the probability Pli, and only took the occurrences\*Pli superior to ‘*maximum – standard deviation’*

Results:

Conditional probabilities applied only on location using the predicted colleges :

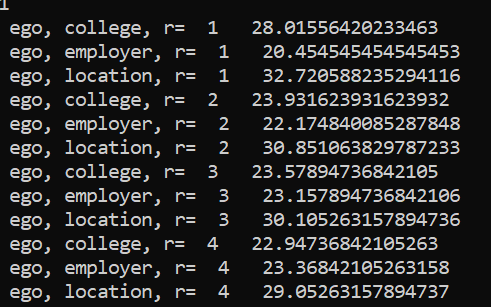


The results are pretty poor, and there is no apparent implementation error, thus leading us to wonder whether this method is accurate. We’ve decided not to use it. Another possibility was to put a coefficient in order to limit the impact of the probability on the predicted result, and do ‘mixed algorithm’ but seeing how poor the results were, we didn’t want to continue on this model, because this model is based on an assumption, that a user goes to a college C1, but seeing the accuracy, the hypothesis is only 41.6% of the time true…

## DIFFERENT LEVELS EGO NETWORK

We wanted to go back to the naive method and extend it using the ego network. There are different levels for the ego network and we can extend the naive method to the level 2, level 3 or level 4 ego network. (level 2 meaning the neighbours of the neighbours, level 3 their neighbours also, etc).

The results we obtained are the following : (r is the radius of the ego network)



We can see the ego network-based method is quite performant, however it is still less than the combination of Louvain and ego network for college and location.

Références Bibliographiques

Annexes

En général, on regroupe toutes les annexes en fin du document.

Annexe 1

* + Annexe 2
    1. Annexe 3

|  |
| --- |
|  |