I - Problem reformulation

1 - Introduction :

As data scientists working at Linkedin, we want to optimize a marketing campaign for a restaurant located in the Bay Area. Our data represents a portion of individuals using Linkedin, their locations, their employers, and their colleges. These datasets are equally incomplete at 60%. That's why, to target influencers that will promote this restaurant, we will have to infer the missing data to make sure we are targeting accurately. We will also need to set a strategy for the influencers' selection: which criteria do we take into account ? and how do we measure our success ?

2 - Problem and objectives definition :

Our main goal is to maximize the campaign's impact by targeting individuals with the highest potential to promote the restaurant among their social circles. But before that, we need to accurately forecast the missing information. In other words, the problem entails predicting the geographical locations of individuals within a social network using a range of features (graph connections, university attended and past employers). Our aim is to enhance the accuracy of our predictions compared to the ground truth data that we are provided with .

3 - Strategy :

To achieve our objectives, we will employ the "birds of a feather flock together" principle presuming that individuals with similar characteristics are geographically close. Additionally we will use machine learning models and try to optimize their parameters. To choose the most influential individuals, we will calculate for each one a score (between 0 and 1) that takes into account many characteristics that will be detailed in the resolution section. The weight of each characteristic will be defined according to its importance after studying the statistics of the graph.

4 - Qualification of success :

Success will be measured by the accuracy of our data prediction model and the effectiveness of our influencer identification process. A "good" accuracy in our case would be one that enables effective targeting of individuals within the Bay Area for the marketing campaign (and by “effective targeting” we mean having similar results of top 10 influencers comparing when them with the ground truth), taking into account business objectives, data quality (especially biases) and practical considerations (by trying different and relevant algorithms).

II - Graph statistics and attributes distributions

1 - Graph statistics :

To provide a comprehensive understanding of our data, we have calculated statistics about the graph. This step was crucial to have an overview of the graph’s complexity. After writing the appropriate code snippet, we have obtained these results:

```
Graph Statistics:
Number of Nodes: 811
Number of Edges: 1597
Density: 0.004862157677611849
Average Degree: 3.938347718865598
Number of Connected Components: 1
Size of Largest Connected Component: 811
Diameter: 19
```

figure 1 - graph statistics

- **The number of nodes and edges** indicates that the graph is relatively large and complex and also tells that there is a considerable number of connections between nodes. This indicates that there’s potential for a wide reach with the marketing campaign.

- **The density** of the graph is low which suggests that the graph is sparse: only a small fraction of possible connections exist and not all nodes are densely interconnected. That’s why we will have to focus on influence metrics to ensure the campaign reaches a significant portion of the network.

- **The average degree** can be an indication of how much an individual can have an influence: with a degree higher than the average, he probably has more potential

- **The presence of only one connected component** indicates that the network is cohesive and that influence can potentially flow across all nodes. This simplifies the task of identifying influencers as we can focus on the entire network rather than isolated components.

- **Size of largest connected component:** The fact that the largest connected component includes all nodes reinforces the idea of a cohesive network

- **Diameter:** it indicates the maximum distance between any pair of nodes. In our case, we can consider it high compared to 811 nodes which suggests that some nodes are quite distant from each other.

2 - Attributes distribution :

Now we will look into the different attributes of our data to enhance the transparency and reproducibility of our analysis and enable a better interpretation of the results. For each parameter (name, employer, college and location) we will provide a summary statistics and count the occurrences of different attribute values to understand their distributions. We obtained the following results:

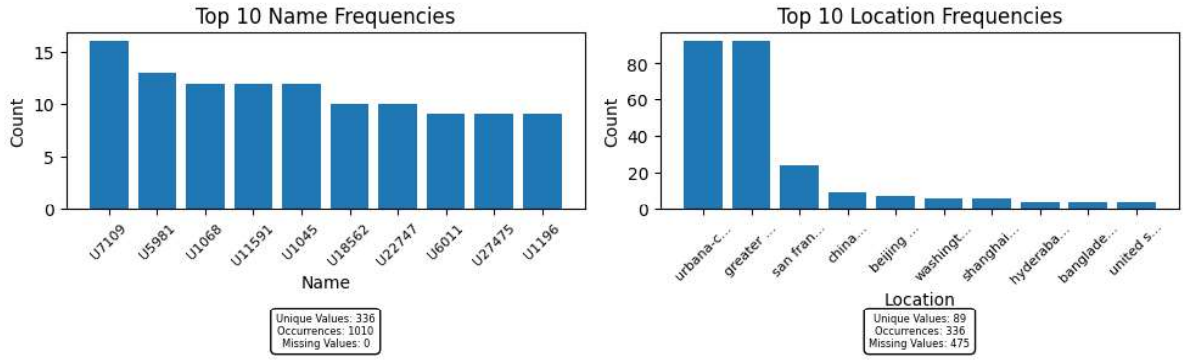

figure 2 - names and locations distributions

Distribution of names: we can see that among the total number of occurrences there is less than half of it as unique values. This proves the presence of duplicate entries: a node can have multiple entries for each attribute. This leads us to question ourselves: if an individual has multiple employers, locations or colleges, which one do we consider? In our case, as we don't have any reference of timeline, the safest way is to keep all the data but we should take into account that this will affect the quality of the results.

Distribution of locations: there are 89 unique locations which indicates the diversity of geographical locations. We also have 336 occurrences. So, some locations are more common compared to others. We can see this in the case of the two most frequent locations having each a count of over 80. This overrepresentation may introduce bias to our analysis if we consider the highly represented locations.

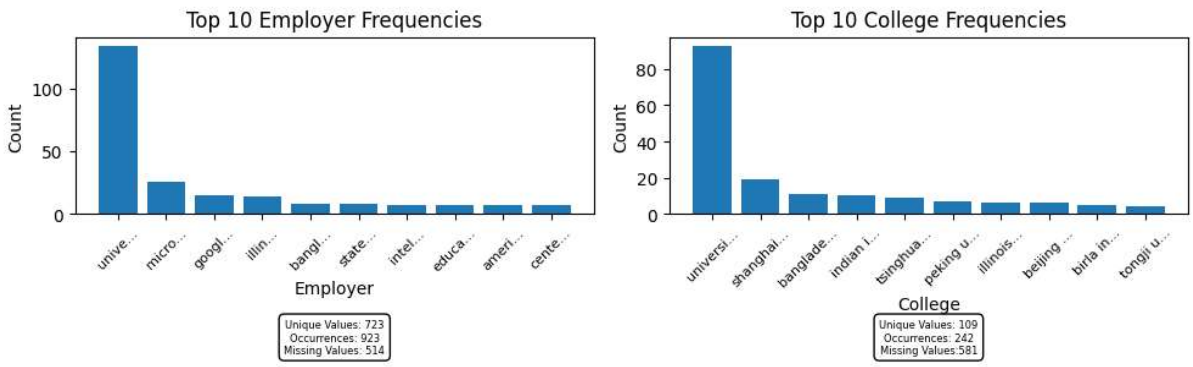

figure 3 - employers and colleges distributions

The same problem is valid for the employer and the college distributions. The most frequent employer and the most frequent college are highly represented compared to other employers and colleges among the top 10. So, it's essential to approach our analysis with awareness of the potential biases and limitations associated with such concentration.

Another notable observation is that all of the nodes for which we have information about their employers also have location data available. This raises the question: does employment information provide meaningful insights for our prediction task?

Given that a node without location information also lacks employment details, it seems that there may be limited utility in leveraging employment data alone for predicting node locations. Since nodes are not interconnected based on this feature—nodes without location information cannot be linked through employment connections—it may not contribute significantly to our predictive model.

3 - Influence parameters definition :

In this section, our goal is to determine how to measure the person's ability to influence people. As suggested in the word " influencer", the aim of such individuals is to promote something in order to reach the mass audience. Knowing that in our case each node represents an individual and the edges connecting theses nodes denote relationships, two approaches are valid:

- a node that has a maximum number of connections has potential to influence these individuals immediately
- a node that can reach a maximum number of nodes throughout a route of nodes that are interconnected can have a major influence

To determine which approach is more accurate in our case, we consider different centrality measures ("centrality" is essentially a measure of the importance of a node in a graph)

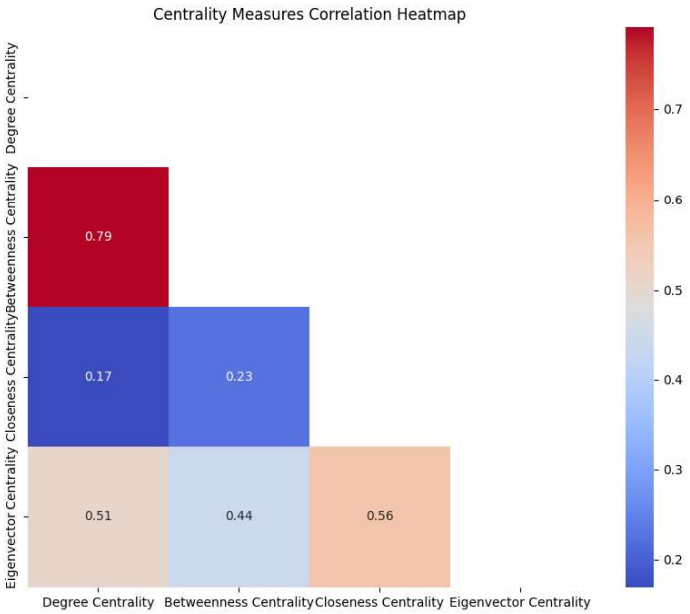

figure 4 - correlation between centrality measures

The figure above shows that the highest correlation is between the degree centrality (measures the number of direct connections) and the betweenness centrality (quantifies the extent to which a node lies on the shortest paths between pairs of other nodes). This suggests that highly connected nodes play a crucial role in mediating interactions and influencing the flow of information

III - Problem resolution

After conducting a thorough analysis of the available data, which includes information on locations, employers, colleges, and nodes, we reiterate our objective : Our primary aim is to maximize the impact of our campaign by targeting individuals with the highest potential to promote the restaurant within their social circles within the Bay Area . Before achieving this goal, accurately forecasting missing information is paramount.

Essentially, our challenge revolves around predicting the geographical locations of individuals within a social network using various features, such as graph connections, universities attended, and past employers. Our goal is to enhance the accuracy of our predictions compared to the provided ground truth data.

1 - Nodes Location :

a. Supervised Machine Learning Approach :

An initial approach we considered was to use nodes with known locations as the training dataset for a supervised machine learning model, which could then be applied to nodes lacking location information. We began by employing the AdaBoost algorithm on nodes, leveraging employment and college information. However, due to the absence of location

data for nodes without employers, this approach resulted in a very low accuracy. Furthermore, we failed to use the valuable connection information provided by the LinkedIn graph.

Recognizing the potential of leveraging graph information to enhance model performance, we sought to identify which graph-derived features are crucial and how to represent them effectively in the training dataset.

The first pertinent information we identified is the connections between nodes, as LinkedIn connections can reflect common professional interests, similar professional experiences, shared academic backgrounds, or common career interests. Individuals may have worked together in the past or attended the same university, which could ultimately lead to sharing the same location.

The presence of LinkedIn connections serves as a valuable feature for predicting node locations. By leveraging the connections between nodes, we can uncover underlying relationships that may indicate shared geographical proximity. For example, if Node A and Node B are connected on LinkedIn, and Node A's location is known, it's likely that Node B's location is also in close proximity. Therefore, incorporating LinkedIn connections into our predictive model can enhance its accuracy by inferring location information based on social connections.

Additionally, we considered the importance of neighborhood location information. If the location of a node's neighbor is known, it provides valuable context for predicting the node's own location. For instance, if Node A is connected to Node B on LinkedIn and Node B's location is known to be in a specific neighborhood, it increases the likelihood that Node A's location is also within or near that neighborhood. Therefore, incorporating neighborhood location information of neighboring nodes can further improve the precision of our location prediction model.

The concrete steps :

1. The nodes with known locations were merged with those containing employer and college information based on the node's name. This consolidation enabled us to have all relevant information for nodes with known locations in one dataset.

2. Utilizing the “ LabelEncoder ”, we encoded the location, employer, and college information from strings to integers. This transformation is vital for the machine learning algorithm, as it requires numerical inputs for processing.

3. To represent the graph in our data, we constructed a matrix where rows represented the nodes to be processed, and columns represented all nodes in the graph. If an

edge existed between two nodes, a value of 1 was placed in the corresponding cell of the matrix; otherwise, it was filled with 0.

4. To incorporate neighbor locations, Another matrix was created, with rows representing the nodes to be processed and columns representing nodes with known locations. If an edge existed between a node in the row and a node in the column, the encoded location value was placed in the corresponding cell, otherwise, all cells were initialized to -1.

5. All matrices and pandas dataframes were concatenated to create a comprehensive dataset.

We followed this approach to create the training dataset, then applied the AdaBoost algorithm to train our model. Subsequently, we reapplied the same method to prepare the dataset for nodes with unknown locations and provided it to AdaBoost for generating predictions.

We obtained the following accuracy : 

Note : *All the given accuracies incorporate all nodes without location in the dataset, including those without location predictions provided by the algorithm. This ensures that nodes lacking location assignments from the algorithm are also considered in the evaluation process*

Concerned about a potential overfitting of the machine learning algorithm to the training dataset, we opted to reduce the dataset by removing 20% of the total training data. Interestingly, subsequent observations revealed that the model predicted a total of 5 distinct locations instead of the expected 4, sparking optimism regarding the accuracy of the predictions. However, despite this shift in prediction diversity, the accuracy metric remained unchanged. This outcome suggests that although the model diversified its predictions, it did not necessarily improve in terms of overall accuracy.

b. Unweighted Louvain Algorithm

The Louvain Algorithm is a community detection algorithm designed to identify groups of nodes with dense connections within a graph.
We select the Louvain Algorithm for its effectiveness in identifying cohesive communities within a network, which can reveal underlying patterns of shared characteristics among nodes. By applying this algorithm, we can capture the inherent structure of the network and leverage it to predict node locations based on community membership.

In contrast to Methodology 1, where location, employer, and college information are directly used for prediction, Methodology 2 integrates these attributes into the graph structure. This

ensures that connections between nodes are preserved while incorporating additional information.

The concrete steps :

1. In this methodology utilizing the Louvain Algorithm, we start by adding the graph with node attributes such as location, employer, and college.

2. With the augmented graph containing both connection information and node attributes, we apply the Louvain Algorithm to detect communities.

3. We assume that nodes within the same community are likely to share similar attributes, including location, we assign a single location to all nodes within each community. We predicted the location of nodes within each community using a majority vote approach. This involves determining the most frequent location within each community and assigning it as the predicted location for all nodes within that community.

We obtained the following accuracy : 

Opting for the majority vote algorithm in location prediction poses some challenges. We encounter two predominant location values in the dataset, and these two values tend to dominate the voting process. Consequently, predictions made by the majority vote method may be biased towards the overrepresented locations. In an attempt to address this issue, we tried removing certain nodes associated with the two overrepresented locations. However, this led to disastrous results, with a significantly low accuracy. Therefore, we decided to stick with our original algorithm without removing nodes associated with overrepresented locations.

c. Weighted Louvain Algorithm

In this approach, we revisit the Louvain algorithm methodology but introduce edge weights to enhance its performance. Adding weights to edges allows us to capture the strength of connections between nodes more effectively. A higher weight indicates a stronger relationship or shared characteristics between nodes, which can significantly improve the accuracy of our predictions.

The weight between two nodes in the graph is determined by three factors:

- **Existence of an Edge** : If there is a direct edge between two nodes, we assign a base weight to capture the presence of any connection.
- **Shared College** : If the nodes share a common college, we increase the edge weight to reflect this shared attribute.
- **Common Employers** : If the nodes have one or more common employers, we assign a weight proportional to the number of shared employers

Once the weighted graph is constructed, we apply the Louvain algorithm to detect communities within the network. This algorithm takes into account the weighted edges, allowing for a more nuanced analysis of community structure based on the strength of connections between nodes. Subsequently, we employ the majority voting approach within each community to determine the most prevalent location, thus assigning the same location to all nodes within the community.

The concrete steps :

1. In this methodology, we used the graph with the added node attributes such as location, employer, and college.

2. We assigned weights to the edges of this graph based on the degree of similarity or shared characteristics between connected nodes. These weights are determined considering factors such as shared colleges, common employers, and the number of common employers.

3. We apply the Louvain Algorithm to detect communities on the constructed weighted graph.

4. We apply the same principle as in step 3 for the unweighted graph to predict the locations.

We ended up with the following accuracy : `Accuracy: 0.2168421052631579`

Why does the unweighted graph show better performance than the weighted graph ?

The partial availability of data leads to a significant loss of information regarding other employers and colleges not included in the dataset. Additionally, if a feature such as employer or college is overrepresented in the data, the Louvain algorithm may tend to form communities around these features. Consequently, this tendency can result in artificial clustering of nodes solely due to the presence of certain features and the weight assigned to them. Although the Louvain algorithm is robust in many cases, it can be sensitive to bias in the majority data, which can affect how communities are detected in the network.

2 - Selection of top 5 influencers :

After preparing and completing the missing data, we will select the top influencers for this marketing campaign. Many criteria should be taken into account to make sure that our selection is accurate. That's why we will attribute a score for each individual using a criteria weighting approach.

First of all, let us define our criteria: As we have seen earlier, the influence of an individual can be measured differently depending on the goal of our targeting.

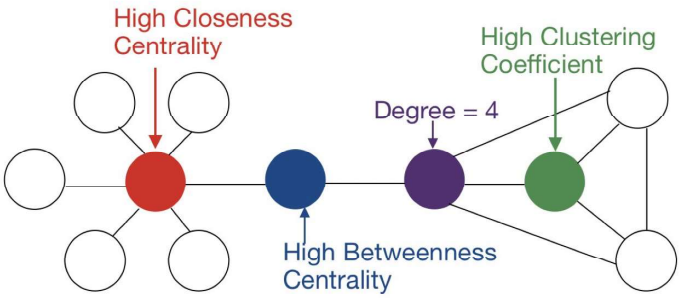

figure 5 - different characteristics of nodes

The figure above shows the difference between each measure and how each one can qualify the influence of a node in a different way. So, we decided to combine some of them depending on their relevance to not “discriminate” some influencers in the favor of others by choosing only one criteria.

- **Betweenness Centrality**: Since we have chosen to use betweenness centrality as a measure of influence, we assign it a relatively high weight to reflect its importance in determining influencers: The higher the betweenness centrality, the more important the node is in the network as an intermediary between other nodes.

- **Degree** : A node's total number of connections can be an indicator of its influence potential. However, we don't want to overweight this criterion in relation to betweenness centrality, so we assign it a lower weight.

- **Number of Bay Area Connections:** As our target is the Bay Area, the number of the influencer’s connections in this region can be an important criterion. We therefore assign it a significant weight.

- **Clustering Coefficient:** A node's clustering coefficient measures the extent to which its neighbors are connected. This can be an indicator of a node's importance in its own local cluster. We give it a relatively low weight.

→ This analysis leads us to the following choice of weights:

```
# Définir les poids pour chaque critère
weight_num_bay_area_connections = 0.3
weight_degree = 0.2
weight_betweenness = 0.4
weight_clustering = 0.1
```

figure 6 - weights of parameters

We have also normalized the obtained value for each criteria to make the values in the score comparable. Finally, we calculated the score of each individual using the data we obtained with the best accuracy and selected the top 10 influencers:

```
Node: U27287 | Influence Score: 0.9023438558460913
Node: U7024 | Influence Score: 0.49446648954321754
Node: U8670 | Influence Score: 0.28473831487042334
Node: U7091 | Influence Score: 0.23833273625846935
Node: U11566 | Influence Score: 0.23409832789823906
Node: U11591 | Influence Score: 0.22612588015534799
Node: U4562 | Influence Score: 0.21264381947863012
Node: U1045 | Influence Score: 0.20556045793458916
Node: U4661 | Influence Score: 0.20430650003160358
Node: U22747 | Influence Score: 0.19237262877104358
```

Top 10 influencers using the louvain algorithm on unweighted graph predictions

```
Node: U27287 | Influence Score: 0.9023438558460913
Node: U7024 | Influence Score: 0.660830125906854
Node: U8670 | Influence Score: 0.25201104214315057
Node: U11566 | Influence Score: 0.2395528733527845
Node: U7091 | Influence Score: 0.23833273625846935
Node: U4562 | Influence Score: 0.23718927402408468
Node: U11591 | Influence Score: 0.22612588015534799
Node: U1045 | Influence Score: 0.2191968215709528
Node: U4485 | Influence Score: 0.1997175718212394
Node: U4661 | Influence Score: 0.18794286366796722
```

Top 10 influencers among influencers of the ground truth data

IV - Conclusion

The results present a challenging scenario, as the achieved accuracy falls short of our expectations. It suggests that alternative algorithms, which we are not yet familiar with, might yield better results.

Also, the quality of the data significantly influences the outcomes; we must address the bias inherent in our data, as it represents different attributes with highly variable occurrences. Some attributes repeat frequently compared to others (especially in the case of location: we have 89 different locations, but our predictions always lead to 4 different locations), which can significantly impact the prediction of missing data. This variability in attribute occurrence rates introduces a potential bias into our analysis, potentially skewing the accuracy of our predictions for missing data.

However, even with this consideration, the top 10 influencers exhibit relatively similar scores when we compare our results with those obtained when we calculate the score for the ground truth data. Thus, the impact on influencer selection is not substantial.

Moreover, except for the top two influencers, the scores diminish notably. This raises questions about the necessity of including the Bay Area location as a factor in the analysis.

V - References

[Social Network Analysis: From Graph Theory to Applications with Python
Goldenberg | Towards Data Science]

Application of Graph Theory in Social Media
Social Media Influencer Identification Using Graphs
Paper Title
What are Boosting Algorithms and how they work
Louvain method - Wikipedia