Challenge #1 Report

Challenge 1 - The Illusion of Hidden Data - Graph Theory

Team 1

Antoine HORRER, Kadriye Nur BAKIRCI

Contents

Business Understanding

Problem Statement

Background

Our client, a restaurant in the Bay Area (San Francisco), wants to launch an online marketing campaign to promote his/her/their restaurant. Our colleagues from the marketing team want to identify the most influential people on LinkedIn who can help promote the restaurant. The problem is to identify the top 5 most influential people on LinkedIn who are located in the Bay Area and most likely to promote the restaurant. The challenge is that not all users have provided their location in their profile, and the restaurant wants to target people who are actually located in this location.

Stakeholders

The stakeholders are the restaurant owner/owners, the LinkedIn marketing team, and the targeted LinkedIn users for the campaign.

Risks and Assumptions

In our problem the possible risks could be:

- inaccurate or incomplete data
- bias in the influencer selection process

to solve this specific problem our estimations include the followings:

- the homophily principle, which states that people who are connected on LinkedIn are likely to have similar characteristics and interests
- influencers located in the Bay Area are more likely to promote the restaurant than those who are not.

Objectives

The main objective is to identify the most relevant 5 influencers who can help promote the restaurant and increase its visibility and customer base.

It is important to distinguish the 5 biggest influencers and the 5 people who influence the most people. Indeed our goal is to reach a maximum of people in the desired region. If we choose the 5 most influential people who influence the same people, the goal will not be reached. It is therefore better to choose an influencer who has a little less impact but in another sample of the population.

The specific objectives are to:

- Preprocess datasets which are named “employer”, “location” and “college” and make analyze LinkedIn graph data.
- Develop strategies to fill in missing data using the attributes of the user's neighbors.
- Evaluate the model by comparing predicted locations to ground truth.
- Deploy the model if it performs well, or refine and re-evaluate the model until satisfying results are obtained.

- Use network analysis techniques to identify the most 5 influential people who have an impact on diverse popularity on the network.
- Select only those who are located in the Bay Area and people who leave other areas but have connections with the people who live in this area.

Problem Translation

From a business problem to data analysis

Our business problem: A restaurant in the Bay Area (San Francisco) wants to run an online marketing campaign by identifying influential people on LinkedIn who can help promote his/her/their business.

Translation into the data analysis problem: The data analysis problem is to identify the top 5 most influential nodes in a LinkedIn graph, where each node corresponds to a user and each edge corresponds to a connection between users. The problem can be broken down into the following tasks:

1. Do data analysis to understand our data and do some graph analysis to get information about nodes and edges for example, node attribute is the user's profile information (e.g., name, age, gender), their location (if available), and other relevant features and edge attribute is the strength of the connection between the users (e.g., number of interactions, frequency of communication).
2. Identify nodes that are missing information, develop strategies and use the homophily principle to fill in the missing information.
3. Use graph theory algorithms to identify the most influential nodes in the network.
4. Filter the influential nodes based on their location attribute. We only keep the nodes whose location attribute is in the Bay Area. Also, we will check the people who live in a different area but have connections in this area.
5. Evaluate the performance of our model by comparing the predicted location of the users with their actual location.

Why should we use graph theory?

The use of graph theory and degree centrality measures can be a powerful tool for exploring the relationships between individuals and their educational and occupational experiences and locations. Using these tools, one can identify the most influential nodes in the network and determine how they affect the rest of the network, which can help to better understand the social and occupational dynamics at work in your data set.

Data Understanding

Fill empty nodes

Dataset Description

In our problem, we have 4 main datasets and 3 ground-truth datasets. Four main datasets include empty, employer, location, and college. The empty dataset includes the node that they have missing information about the employer, location, and college. On the other hand employer, location, and college datasets include both nodes which are different from the empty dataset and information about employer, location, and college respectively. Three ground-truth datasets are mainly employer, location, and college to evaluate our strategies after filling empty nodes.

Dataset	**empty**	**employer**		**location**		**college**	
Attributes	name	name	employer	name	location	name	college
Count	475	923	923	336	336	242	242
Unique	475	297	723	336	89	230	109
Top	U27476	U5981	university of illinois at urbana-champaign	U1313	urbana-champaign illinois area	U1045	university of illinois at urbana-champaign
Frequency	1	13	76	1	92	3	66
Dataset Shape	(336,2)	(923,2)		(336,2)		(242,2)	

In the above table, we can see the statistical description of each dataset. Based on these statistics we can say that the employer dataset has 923 values and 297 of them are unique, the college dataset has 242 values but 230 of them are unique. That's why we combine similar nodes into one row because the order of employer or college is not important. We don't have any null values, and data types are perfectly defined that's why we don't need to do further analysis.

Graph Description

Our network summary is:

Number of nodes: 811

Number of edges: 1597

Average degree: 3.938347718865598

And in our network:

514 nodes have no employer attributes among the 811 users in the graph

581 nodes have no college attributes among the 811 users in the graph

475 nodes have no location attributes among the 811 users in the graph

This is the obvious result because from the table, for example, we can see the employer dataset has 297 unique values, and since the network has a total of 881 nodes we have 514 nodes that have missing information. To fill up this information we developed several strategies but before that homophily test is done which is the principle that contact between similar people occurs at a higher rate than among dissimilar people.

Test results are:

Assortativity coefficient for employer attribute: 0.019418389262021312

Assortativity coefficient for location attribute: 0.04932742332337936

Assortativity coefficient for college attribute: 0.04848618998430686

Which are pretty bad which means there is no similarity between nodes for each attribute.

Strategies

The first strategy is to calculate the Jaccard coefficient which measures the similarity between two sets. Here, we measure the Jaccard coefficient between the set of neighbors of a node with known attributes and the set of neighbors of the node with the missing attribute. The assumption is that two nodes that have similar neighbors are likely to have similar attributes. We then choose the most frequently used attribute value among the similar nodes.

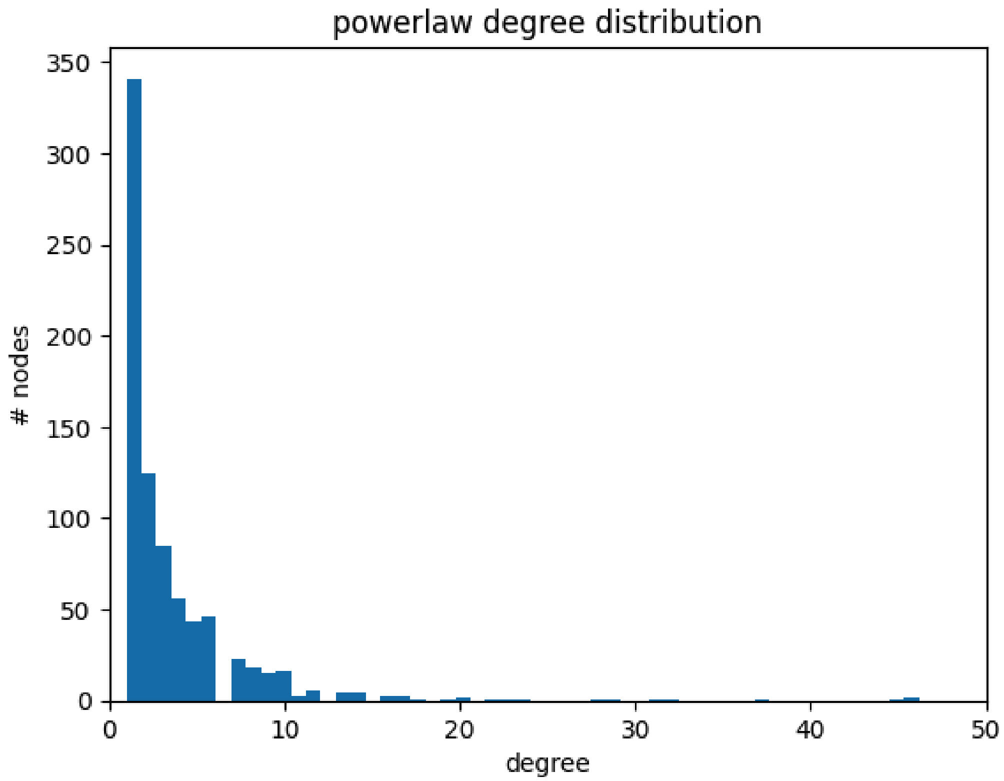
The second strategy is the weighted relational classifier. The assumption is that two connected nodes are likely to share the same attribute value. Here we take into account the frequency of the attribute values among the neighbors and the overall distribution of the attribute values in the dataset.

The third strategy is to predict the missing attribute using the homophily principle, which states that nodes that are connected in a network are more likely to share the same attribute value. This strategy separates into two parts. The main difference between the two parts is that the second part can handle multiple attribute values for a given node, while the first part assumes that each node has only one attribute value.

Find influencers

The challenge we faced was determining the top five influencers for a restaurant using a graph pulled from LinkedIn profiles. As we mentioned earlier, our goal was to identify the five most effective influencers for a marketing campaign. To achieve this, we needed to find influencers with many connections to other users, thus representing a great influence on the social network. Furthermore, it was essential that the five nodes selected did not share the same "friends", in order to guarantee maximum reach for our marketing campaign. Using sophisticated network analysis techniques, we were able to identify the five most relevant influencers for our clients and provide them with valuable recommendations for their successful marketing campaign.

First of all, we’ve started to plot the bar chart of the degrees of the graph to have an idea of the distribution. We can see that the majority of the nodes are not very related to others but some of them have still linked to a lot of neighbors.



Afterward, we calculated how many neighbors each node had by ranking them and selecting only the 10 with the most neighbors.

```
The 5 nodes with the highest degree are: [('U7972', 46), ('U8670', 46), ('U1045', 58), ('U7024', 74), ('U27287', 122)]
```

```
The 5 next nodes with the highest degree are: [('U4485', 29), ('U5977', 31), ('U15267', 32), ('U4562', 37), ('U7091', 45)]
```

Now that we know the degree of the largest nodes we want to know how much it is grouped with these neighbors. For this, we will calculate the grouping coefficient. We want to have the lowest possible coefficient so that the node is the least close to these neighbors.

```
{'U27287': 0.023438558460913157,

 'U7024': 0.0011106997408367272,
 'U1045': 0.011494252873563218,
 'U8670': 0.035748792270531404,
 'U7972': 0.02995169082125604}
{'U7091': 0.008080808080808,
 'U4562': 0.08258258258258258,
 'U15267': 0.17540322580645162,
 'U5977': 0.15913978494623657,
 'U4485': 0.07389162561576355}
```

The node U7091 which do not take part to the 5 biggest nodes seems to be interesting having a very low clustering measure.

To check if our result is good we will use the PageRank algorithm and the relevance calculation.

The calculation of the relevance score of each page is based on two main factors: the number of inbound links pointing to the page and the relevance of the pages pointing to this page. Pages that have many inbound links from relevant and popular pages get a higher relevance score than those with few links or low-quality links.

This algorithm was created by google and can be used to find out if our influencers are themselves followed by people with a certain notoriety, this can be useful in our marketing campaign.

This is our result:

```
1. U7024 avec un score de PageRank de 0.0371
2. U27287 avec un score de PageRank de 0.0345
3. U1045 avec un score de PageRank de 0.0257
4. U7091 avec un score de PageRank de 0.0214
5. U7972 avec un score de PageRank de 0.0168
6. U8670 avec un score de PageRank de 0.0138
7. U22747 avec un score de PageRank de 0.0109
```

```
8. U4562 avec un score de PageRank de 0.0094

9. U4485 avec un score de PageRank de 0.0086

10. U14068 avec un score de PageRank de 0.0070
```

We see that as we said before, the node U7091, even if it has a lower degree than U8670, is more interesting because its neighbors are themselves more followed by important nodes.

Evaluation

Fill empty nodes

Among all the strategies that we used to fill in missing information in an empty dataset, the best results are given by the homophily method part two since it can handle multiple attribute values for a given node. But still, it was not a very good result because in our opinion the homophily test results are also pretty bad which means there is no similarity. To improve this strategy it could combine with other strategies like first calculating the Jaccard coefficient or other similarity techniques then implementing homophily, or incorporating information from other attributes. For example, if we are predicting a missing location attribute, we could also use the employer or college attributes as additional features to predict the location. This would allow us to capture more complex relationships between the attributes and improve the overall accuracy of the predictions. Or maybe it is another possibility to improve the accuracy of the predictions is using machine learning algorithms like the random forest.

Find influencers

Finally, the two methods to find out who were the best influencers, i.e. which nodes had the most neighbors and were as far away from a cluster as possible, gave the same result. The 5 nodes are unquestionably: U7024, 27287, U1045, U7091, U7972. These are the ones that should be contacted to do a marketing campaign on Linkedin to promote the restaurant.