Graph theory : Challenge 1

Group : Daniela Batalha and Ishita Chaudhary

Teachers : Cécile Bothorel ; Yannis Haralambous ; Nicolas Jullien

SUMMARY

1. Business understand

In challenge 1 we are working as data scientist at LinkedIn interested to identify five influences to promote an online marketing campaign for a LinkedIn client, a restaurant in San Francisco Bay Area. The objective is to find the 5 most influential people on the network who would be best to promote the restaurant. For that, we have to maximize the restaurant's reach by selecting the five most influential people in SF Bay Area.

To do this, we formulate the above problem in the form of a graph. We define the LinkedIn graph G= (V, E), where V is the set of nodes; a node is defined as a person, and E (the set of edges) is a subset of VxV, where an edge between the nodes A and B exists in graph G if A and B are connected on LinkedIn. First, we fill the empty location information , and after we identify the five influences by calculating distance coefficients, such as degree, betweenness, centrality and eigenvalue.

2. Data understanding

The data has been extracted from the social network site LinkedIn. We treat a user as an individual node. Each node is provided with employer, college and location as their attributes. The most common drawback of these type of datas is missing attributes, as some people might not fill out their profiles on social media sites.

The total number of users are 811, with 1597 connections between them. The data we have consists of 475 users with missing attributes. Also, the users with known attributes may have more than one employer, college and/or location.
In our data, 297 users have more than one attribute of the same type. We need to complete our data set first in order to analyze it.

3. Data preparation

In a social network, nodes may have attributes, such as gender, location, etc. The property of *assortativity* says that nodes that are connected to each other in a social network tends to be similar[1].

One possible factor that explain *assortativity* is Network homophily. This property says that similar nodes may be more likely to attach to each other than dissimilar ones[1]. We used this concept to fill empty user attributes.
First, we calculated the *assortativity coeficient*, that is, the Pearson correlation between the degrees of pairs of linked nodes. As a result, we obtain -0.22. It means that this LinkedIn network is dissasortative, because the *assortativity* coefficient is negative. The Web, the Internet, food webs, and other biological networks tend to be disassortative [1].

Strategies to fill empty profiles

1. Strategy 0: Base Method

The first strategy to fill empty profiles is the Naive method, where the assumption is that two connected nodes are likely to share the same attribute values. Here, we choose the most frequently used attribute value among the neighbors. Using it, the prediction of location has an accuracy is of 32.35% and 27.39% for employer.

2. Strategy 1: Page Rank

For this strategy, we try to use page rank method of NetworkX library to predict attributes of empty nodes. We sort the nodes in the highest to lowest order of page ranks, and start iterating the list. If the current node with higher page rank has empty attributes, we fill its attributes with base method using strategy 0. Otherwise, we fill the empty neighbours of the node with the attributes of the current node. However, with this method we acheived an accuracy of 21.57% for employer.

3. Strategy 2: Closeness Centrality

We proceed in the same was as strategy 1, with the only difference being that instead of page rank method, we used closeness centrality method (from networkX library) to sort the nodes. The accuracy acheived in this case was same as above. We also tried different methods to sort nodes like betweeness etc., but the accuracy acheived from these two strategies were maximum, but still less than the strategy 0, hence, we decided to stick with strategy 0 for the second part of the challenge.

4. Modeling

Finding influences

The number of connections that someone has is just one possible measure of influence on LinkedIn. Another way to think about influence might be to examine who is often connected with other people. In this analyze, the output include four network centrality measures that are useful in knowing how the information will spread through the network. They are:

Degree

In a social network, the degree of a node defines the total of connections the users have been with. Mathematically, Degree Centrality is defined as below:

$$C_D(p_i) = \sum_{k=1}^{N} a(p_i, p_k)$$

The calculation simply count how many other nodes it's connected to. As a result of the calculation, in challenge 1 we have found the users 'U4568', 'U27661', 'U16141', 'U15272' and 'U22771' as the five users with the highest degree, where, each one present 14, 9, 7,6 and 6 neighbors(connections), respectively.

Betweenness

This measurement of centrality shows which nodes serve as a bridge on the shortest path between other nodes in a network. Mathematically, Betweenness Centrality B(i) of a node *i* in a graph is defined as below:

$$C_B(n_i) = \sum_{j<k} g_{jk}(n_i) \; / \; g_{jk}$$

Nodes with high betweenness may have strong influence within the network because they have an important and strategic position of connection between other nodes. As a result of this calculation, in challenge 1 we have found the users 'U4568', 'U16141', 'U2622', 'U24084', 'U15272' as the five users with location in San Francisco Bay Area with the highest betweenness centrality coefficients, so, these five nodes have the most special position into the network.

Closeness centrality

Closeness centrality is the average length of the shortest path possible from a specific node to all the other nodes in the network. We also understand as the average distance from one node to any other nodes on the network. So, the user with the highest closeness centrality will have the lowest average distance to any other random node and are the nodes that can spread information the fastest.

Mathematically, the Closeness Centrality (Cc) of a node *i* in a graph is defined as below:

$$C_c(p_i) = \frac{N-1}{\sum_{k-1}^{N} d(p_i, p_k)}$$

As a result of this calculation, we have found the users 'U4568', 'U24064', 'U27661', 'U24084', 'U27614' as the five users with the highest betweenness centrality coefficients. So, If we want to send the marketing campaign information to each node in San Francisco Bay Area, we should select these users to transmit it quickly into the network.

Eigenvalue centrality

Eigenvalue centrality is a measure of how influential a certain node is within the network, assigned relative to all the other nodes. It's an algorithm that measures the transitive influence of nodes. A node is high in eigenvector centrality if it's connected to many other nodes who are themselves well-connected. As a result of this calculation in challenge 1, we have found the users 'U4568', 'U27661', 'U24064', 'U27614', 'U27585' as the five most influential users in San Francisco Bay Area, with the highest eigenvalue centrality coefficients.

Page Rank

PageRank computes a ranking of the nodes in the graph G based on the structure of the incoming links. As a result of this calculation, we have found the users 'U4568','U16141', 'U27661', 'U22771' and 'U27614' as the five users with the highest Page rank coefficient.

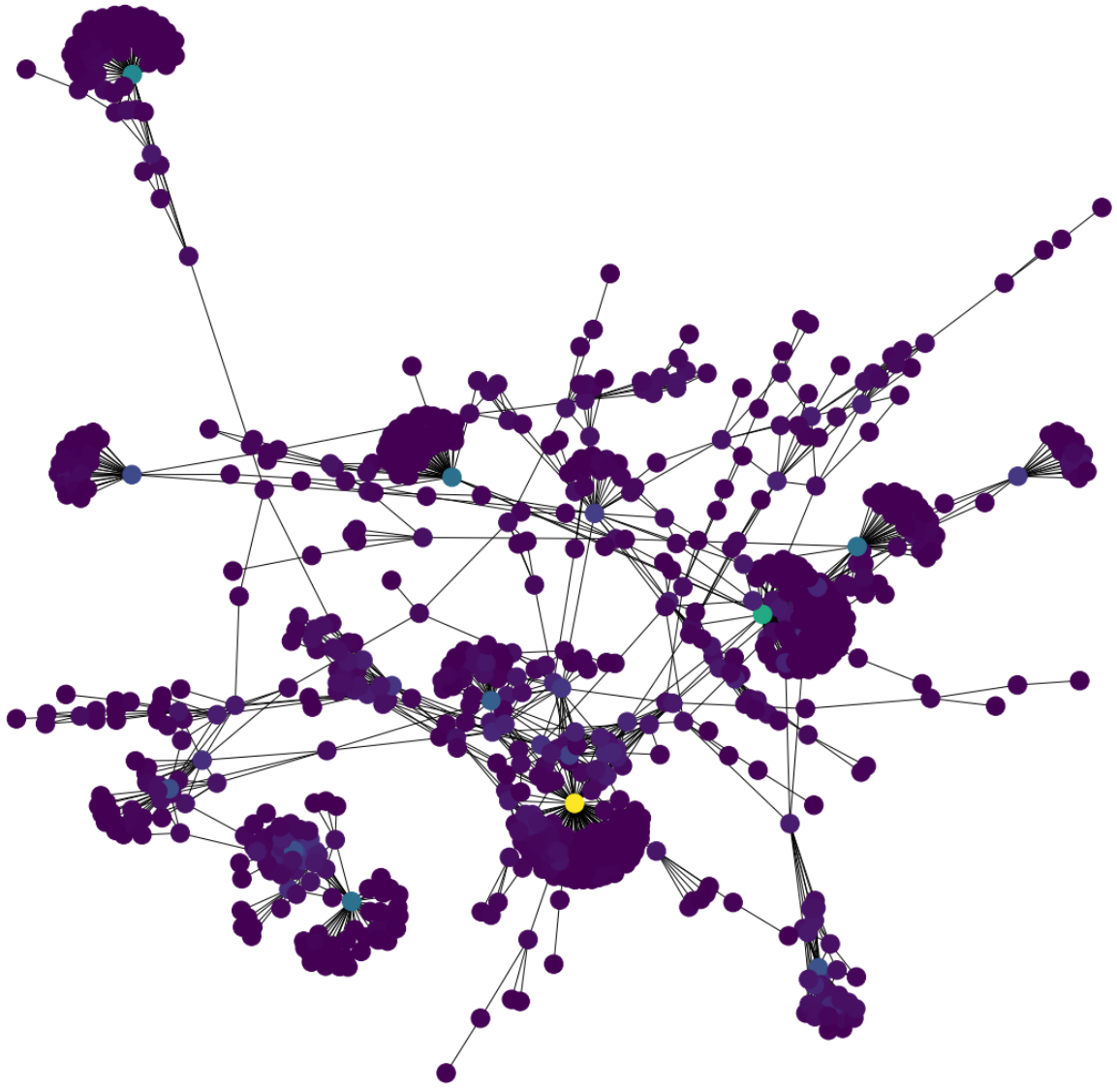
5. Evaluation

The table below summarize the results of users for each centrality algorithm used. We can see that some of them, like 'U4568' and 'U16141' appears several times.

Users \ centrality algorithm	Degree	Betweenness centrality	Closeness centrality	Eigenvalue centrality	Pagerank centrality
User 1	U4568	U4568	U4568	U4568	U4568
User 2	U27661	U2622	U24064	U27661	U16141

User 3	U16141	U16141	U27661	U24064	U27661
User 4	U15272	U15272	U24084	U27614	U22771
User 5	U22771	U24084	U27614	U27585	U27614
Reach	**42**	**32**	**36**	**36**	**41**

We can also visualize the network such that the node color varies with Degree Centrality.



By the graph, we can also see that it's a disassortative network. So, nodes of high degree tend to connect to nodes of lower degree. It's possible to see that we

have alternate vertices with low and high degree. The assortativity degree is -0.22.

6. Conclusion

To conclude, we think that the most appropriate technique to find the five influences for the campaign marketing on LinkedIn will be Degree centrality , therefore the most influence users are 'U4568', 'U27661', 'U16141', 'U15272' and 'U22771' with a total reach of 42.