Facebook Network Analisys PT2

Jorge A. Aque G.

Universidad Politecnica de Yucatan

Social Network Analysis
Merida Yucatan
st1809007

Keywords— Social Network, Facebook, social network analysis, graph, metrics. Independent cascade, maximization influence, threshold model, Facebook Network Analisys.

Abstract

During the Project we will analyze the relationship between different users of the social network Facebook. It is worth mentioning that the positions and / or names of the nodes have been made anonymous for their own privacy.[1]

This Project is the successor of another project in which this Facebook network is analyzed in another way. The fact that this same network was chosen is because we want to seek a full-depth analysis and observe the different behaviors and analyzes that can be obtained from a network. Clearly, another network could be taken for analysis, but at your own discretion it is better to have the same sample of data to analyze and grow in terms of knowledge working with the same example. In the same way, the fact that new things can be obtained on the same network is the most interesting.

It is intended that the purpose of this project part 2 is the analysis of a social network and observe its behavior, and as a main point as a case study, the maximization of information represented in the independent cascade model is sought. During the analysis process it is intended to understand the limitations of the graph, it is that throughout the project the information obtained is the most accurate possible.

I. INTRODUCTION

As we know, being a data engineer involves a lot, really a lot of understanding of the data and you have to be prepared for any type of analysis required for your field of work.

Since the beginning of digital communication, people have looked for a way to be in contact, they have always wanted to be in constant communication regardless of the distance and as we have seen, people do not only have contact with a single person.

This Project consists of resuming the analysis of the Facebook network, so we will see similar processes, instead it lies in the final analysis which is taken as a case study in maximizing the influence between nodes, which consists in obtaining the greatest change in influence on the network in the shortest possible process.

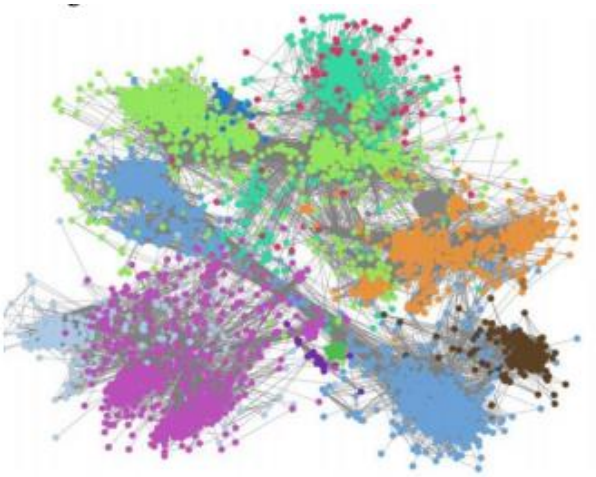
Taking this case study, the process to obtain the visualization will be by modeling it, having as a model the threshold model, which is an effective algorithm due to its low demand for time and resources.

As I mentioned before, the main point of this project is to analyze the network in a different way and as a final point to obtain the analysis and modeling of almost study. Regarding the characteristics of the network, there will not be much change since the information from the previous analysis will be maintained because we want to perform a type of repository of different data analysis on the same network as the main example. Of course, another network could have been taken for analysis, but I repeat, you want to keep the same example in order to be on the same page and that the project is directed to what you are really looking for, in this case, the case study and its modeling.

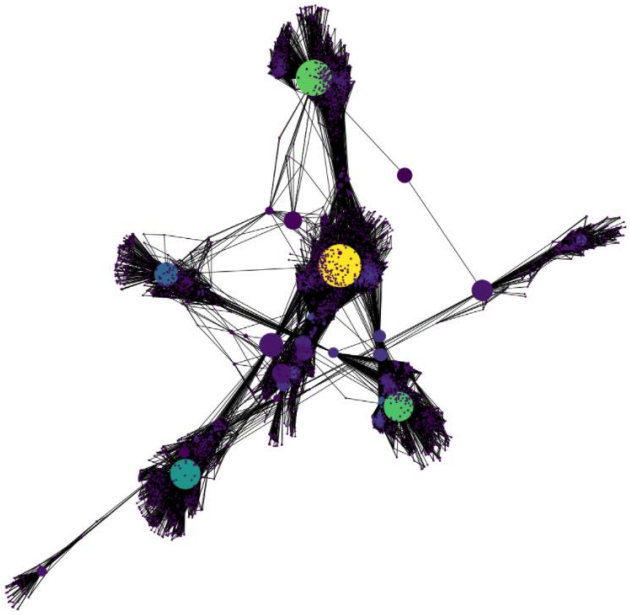
II. Network Characteristics [6]

First of all, the most important step was to take a
Image of the graph to know what type of graph it is and see at a glance what the problems of its analysis could be, in the same way to contemplate how beautiful something as abstract as simply people connected in a social network can be seen.
This visualization had already been shown previously, we can see with the naked eye that there are several small groups in the network.

In the same way, we had previously analyzed that in how many communities formed in the network there were 13, this told us the different groups that existed in the network despite being a connected network, that is, despite the fact that these people were friends and friends in others they formed a different community until there were 13 communities.



During this new analysis I wanted to make a different visualization for this network, as the main objective in this network we sought to observe the nodes by their degree, which we will see below:



As we can see there are several nodes which have great importance, the most notable being the yellow node, this we could say that it is the most important node since it is the one that has the most relationship with other nodes, since the nodes of this social network (which are people) are anonymous, we do not know what type of people this node would be, although we could suppose that it may be someone with political influence since as we know this network has as part of its members some people who belong to the government.

So we know the number of nodes with which we are working is 4039 which is a good number for the analysis. In the same way they follow it, it caught my attention but in the end it makes sense since a person not only has a friend as a single person, what I am talking about is the number of "edges" which I had as number 88234 edges.

As I mentioned before, being able to visualize the network helps us a lot for a quick anailisis and we can see that the Facebook network is a graph that is totally connected which led us to our next metric, the “shortest path”, for this metric the truth I did not know what to expect, the truth is, I thought it would give me a big number like 10, mostly because of the small sections at the ends but in the end as a result I got a shortest path of 4 which surprised me a lot since what it tells us is that if I want to get from person A to person B I have to spend a minimum or rather, on average, between 4 nodes to be able to complete my destination, in the same way it is worth mentioning that not only that I was impressed but also the number that I got the "diameter", as I mentioned before I thought that the number of the metrics were going to be high, in this case I thought that the diameter would be like a 20 but actually I got a diameter of 8 which as we know is the greatest distance between all the pairs points of it.

Finally, the density of our graph as the main idea we must know that we have to obtain a low number for this type of graphs since they are from social networks and I obtained 0.010819963503439287 as a result, As mentioned before, it is possible that this network has this density because it does not have a great connection in all the nodes.

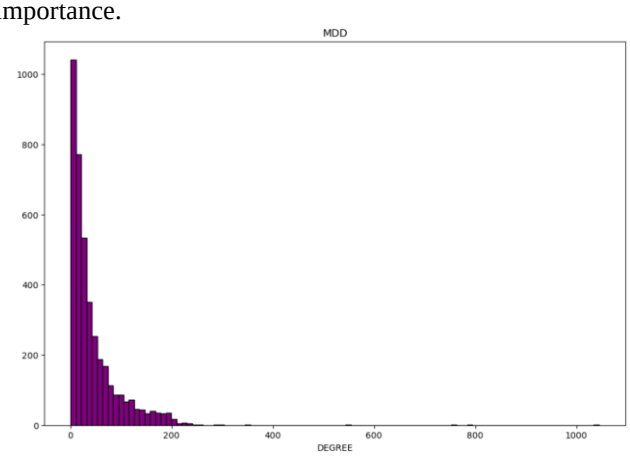
I do not want to talk more about characteristics since I have spoken in the last section about these and what my expectations were, even so this is a small summary of the results obtained.

A. *Summary*

```
Number of nodes:  4039
Number of edges:  88234
Density:  0.010819963503439287
Cluster coefficient:  0.6055467186200876
Diameter:  8
Average shortest path:  3.6925068496963913
Eccentricity:  4
197 nodes are the periphery
1 nodes are the center
```

Characteristics summary

For the last part, as we have already seen regarding the visualization of the degrees of the nodes, we can notice that there are very few nodes that have a high degree, which as we saw in the previous visualization makes sense since very few nodes can be highlighted as nodes with high importance.



Node degree

III. MODEL

A. *The Threshold model*

In mathematical or statistical modeling a threshold model is any model where a threshold value, or set of threshold values, is used to distinguish ranges of values where the behavior predicted by the model varies in some important way. [7]

In other words, the threshold model, like other similar algorithms, looks for a way to change a model using the number of least possible steps, in this case this model was applied to see the influence of nodes on the network, obtaining a number of Minimum iterations, this type of model is very common to use for the analysis of the behavior of groups in a network.
Threshold models are often used to model the behavior of groups, ranging from social insects to herds of animals to human society.

According to Granovetter, the "threshold" is "the number or proportion of others who must make a decision before a given actor does." Which later we will talk about this.

The thresholds of individuals can be influenced by many factors: socioeconomic status, education, age, personality, etc. In addition, Granovetter relates the "threshold" with the utility that is obtained by participating or not in the collective behavior, using the utility function, each individual will calculate their cost and benefit of taking an action. And the situation can change the cost and benefit of the behavior, so the threshold is situation specific. The distribution of the thresholds determines the outcome of the aggregate behavior.

In this case with the future results we can say that the network can be affected in terms of its influence according to the nodes that represent people with political positions, it is like this since as we know this type of people usually have a large amount of influence in society, in this case, on the network.

As we know in our case for the model we use the threshold which will help us to have greater flexibility and expansion for the influence on this network.
In this case the threshold I took for my model was:

Threshold = 0.1

Since it is a common value for this input parameter.
Sure I made several comparisons with different values for this parameter but I decided to let my criteria be the best to get a visualization of the maximum influence.

B. *Case Study*

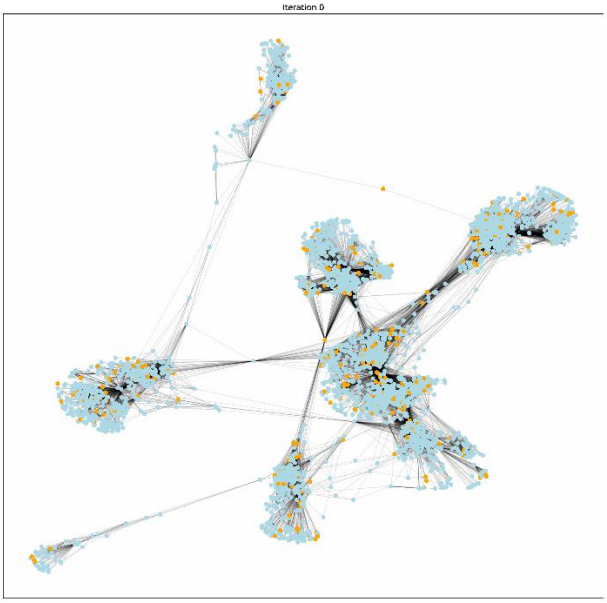
Influence maximization is the problem of finding a small subset of nodes (seed nodes) in a social network that could maximize the spread of influence. In this paper, we study the efficient influence maximization from two complementary directions.

In this case, as I have already mentioned, it is intended that the selected model can show the behavior or the change according to the influence on the network, for this section a gif was obtained as a result to be able to visualize the changes.

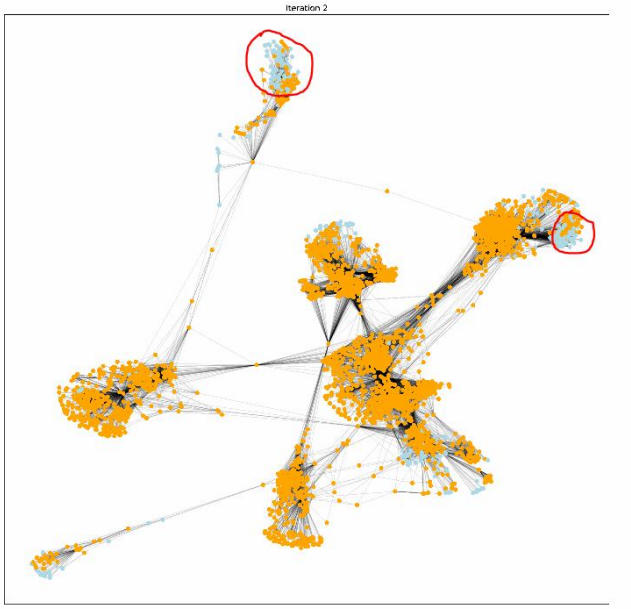
It is worth mentioning that it was tried to use the independent cascade as a model but the execution time was high to such a level that the laptop used could not support it, this is a clear example of the need for code optimization and although there are other ways to implement it, in this case we chose to follow the analysis only in Python.

IV. RESULTS

As I mentioned before, the model for my case study was the threshold model, which I obtained as a result of this small visualization that consists of 4 iterations, which means that in 4 movements the network can be totally influenced by taking certain important nodes. Next I will show how the network is according to iteration number 0, that is, the beginning. As we can see very few nodes are influenced.

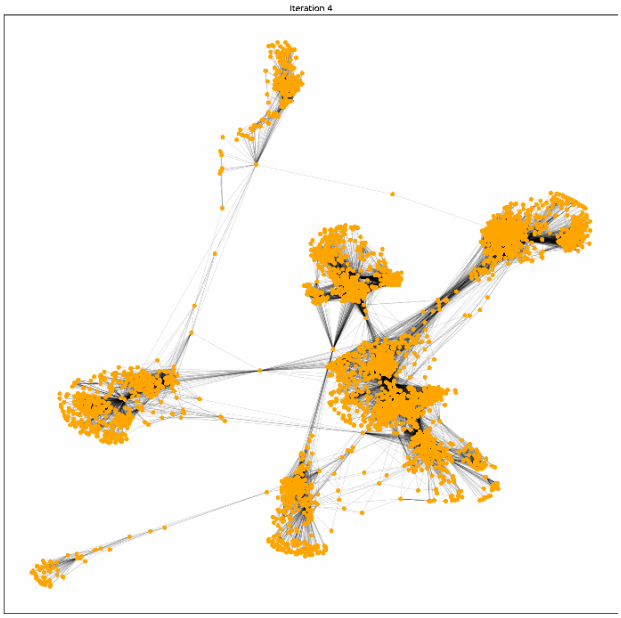


Iteration 0



Iteration 2

For the second iteration we can see that most of the network is influenced, there are still some nodes, for example at the top we can still see some blue nodes which have not been affected yet.



Iteration 4

As we can see, the maximization of influence was completed in the fourth iteration, which is a good result since, as we know, we seek maximization and the fewer movements the better.

In this section it was very helpful to know how to handle the threshold model since as we see the result was excellent, this model was chosen since its computational complexity is better than others, the independent cascade was a model that was taken as the main option but The execution time was very high and I had to resort to other execution modes for the model, so I obtained from this model that despite taking

about 1 minute to make the calculations, it was much better in how much time, this taking it at the discretion own self.

So, as I mentioned before, taking the threshold parameter of 0.1 as input, this was the best result I got with a minimum of 4 iterations.

V. Conclusion

For this Project, as a final conclusion, I found it to be of great help to reinforce past knowledge, such as the visualization of the network, it was something that I really liked since it was very different from the result I obtained in the previous project, I had already mentioned it but I wanted to return to this Facebook network since it seemed a little more like implementing my knowledge since I had already worked with it, the truth is that these types of projects are important since although they are sometimes complex they help us to get used to that we will always find complexity in any analysis, no matter how small. The reason why I used the influence maximization as a case study and the threshold model was by my own criteria since in my opinion the case study seems interesting, it knows how the network can be affected with only 4 iterations and in the same way I think the threshold model is one of the simplest and most important to use for group behavior analysis. I was satisfied with the results, although a much more pleasant way of having this visualization would be to combine the visualization where the degree of each node is noted and its change according to the iterations.

VI. References

[1] Leskovec, J., 2021. SNAP: Network datasets: Social circles. [online] Snap.stanford.edu. Available at: [Accessed 23 June 2021]

[2] "Reading and writing graphs — NetworkX 2.5 documentation", Networkx.org, 2021. [Online]. Available: https://networkx.org/documentation/stable/reference/readwrite/index.html. [Accessed: 25- Jun- 2021].

[3] I. Despot, "Understanding Community Detection Algorithms with Python NetworkX", Memgraph.com, 2021. [Online]. Available: https://memgraph.com/blog/community_detection-algorithms_with_python_networkx. [Accessed: 25- Jun- 2021].

[4] "networkx.algorithms.community.quality.modularity — NetworkX 2.5 documentation", Networkx.org, 2021. [Online]. Available: https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.quality.modularity.html. [Accessed: 25- Jun- 2021].

[5] Klein, Douglas. (2010). Centrality measure in graphs. Journal of Mathematical Chemistry. 47. 1209-1223. 10.1007/s10910-009-9635-0.

[6] Aque (2021). Facebook Network Analysis. GitHub. https://github.com/AqueJorge/Projects/blob/master/Facebook%20Network%20Analisys(Community%20Detection).pdf [Accessed: 25 – Jul -2021]

[7]"Threshold model - Wikipedia", En.wikipedia.org, 2021. [Online]. Available: https://en.wikipedia.org/wiki/Threshold_model. [Accessed: 27- Jul- 2021].

[8] M. Nandi, "Social network analysis with NetworkX", Data Science Blog by Domino, 2021. [Online]. Available: https://blog.dominodatalab.com/social-network-analysis-with-networkx/. [Accessed: 27- Jul- 2021].