LINMA2472 - Algorithms in data science

Final Project: Diffusion of information along Networks

Authors : Heroufosse Gauthier
Degives Nicolas
Gandjeto Martine

2021 - 2022

1 Introduction

The Internet has provided an unprecedented access to information. Therefore, a news spreads today much faster than in the past through television, newspapers or radio. The social networks participate greatly in the access to news. This revolution allows a direct access to information but this access is also amplified for fake news. In order to appreciate the spread of information that we can observe nowadays, we will simulate the spread of a news (false or true) through different networks similar to what we can find online in order to compare their behavior.

Twitter, Facebook, Instagram, ... Each social network works differently. And each one allows to spread information. In order to compare the propagation of this information through these different networks, we will simulate the propagation of a news by different mode of diffusion through two different networks. Our attention is mainly focused on Facebook and Twitter as they are two well known social networks.

2 The Data sets

The two compared networks are built using two data sets from the Stanford network database. The first one contains friends list from Facebook and contains about four thousands nodes. The other one contain followers list with seven thousands nodes. Table 1 below shows the main characteristics of these two networks.

Table 1: Main characteristics of the networks

Network	1	2
Social Networl	Facabook	Twitter
Type	Friends list	Followers list
Nodes	4039	7115
Edges	88234	103689

The reason why it is necessary to use two different networks is that the type of edges is different. Facebook works with a system of friends. When two nodes are friends, they can share information with each other. The propagation can therefore be done in both directions. The edges are then said to be undirected. On the other hand, Twitter works with followers. Only the followed node can send information to the nodes that follow it. In this case, the edges are said to be directed. This difference drastically changes the way the network works and, by extension, the way information propagates within it.

Both networks are a bit too large and require a lot of computing time. So we reduced them by removing some nodes. After reduction, the Facebook network is composed of 285 nodes and 1747 edges and the twitter network of 223 nodes and 1691 edges. It also allows to work with two networks of similar size.

Here is how the two reduced networks looks like :

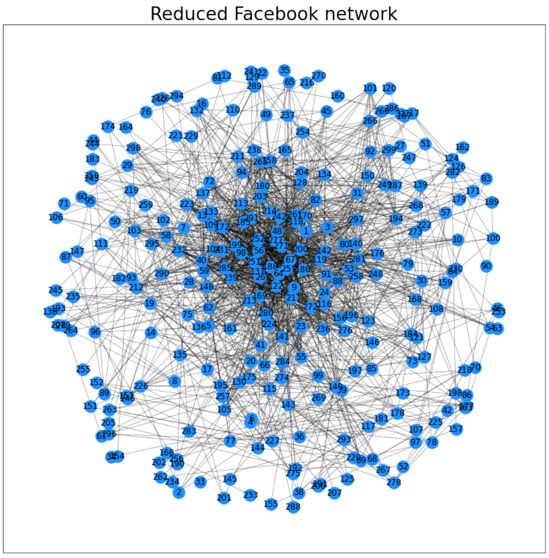


Figure 1: Facebook Network

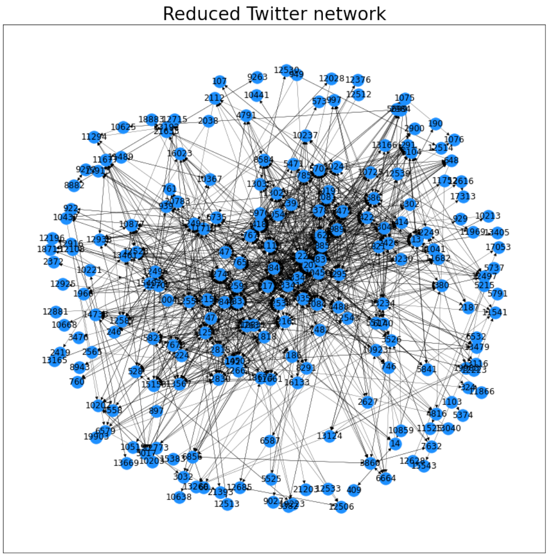


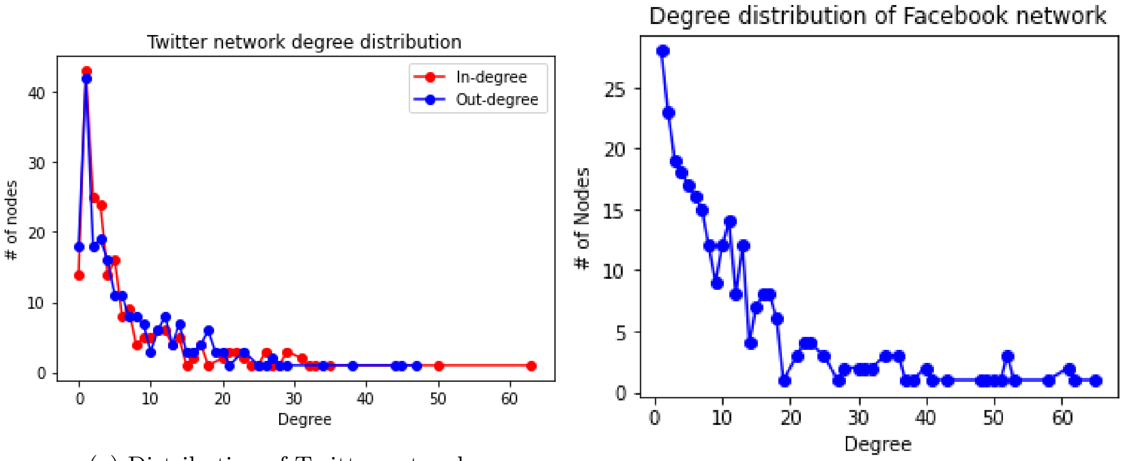
Figure 2: Twitter Network

3 Properties of the two networks

Before starting to compare the diffusion process, we need to be sure that our networks present similar properties.

3.1 Degree distribution and degree assortativity

A property of the full-scale structure of a network that is typically investigated is the distribution of the network node degrees. The degree distribution of the real networks is asymmetric. Directed networks like twitter dataset, have two different degree distributions, the in-degree and the out-degree distributions (between connections where the node is the source/follower and where the other node is the destination/followee). In general, this network is primarily made up of many followers with few connections and few followee with larger degrees. Distributions like this are common for real-world networks. In undirected networks like facebook, this is simply the number of connections a node has. (1)



(a) Distribution of Twitter network

(b) Distribution of Facebook network

Figure 3: Distribution of Twitter and Facebook networks

As we can see, the distribution are quite standard. If we compare this with a Barabasi-Albert graph, we see that the shape of the distribution is similar.

About degree assortativity, we get positive values for both networks. It means a tendency for nodes of similar degrees to connect to each other. It is quite uncommon for this kind of networks, as the Barabasi-Albert model present a slightly negative degree assortativity, meaning that a few

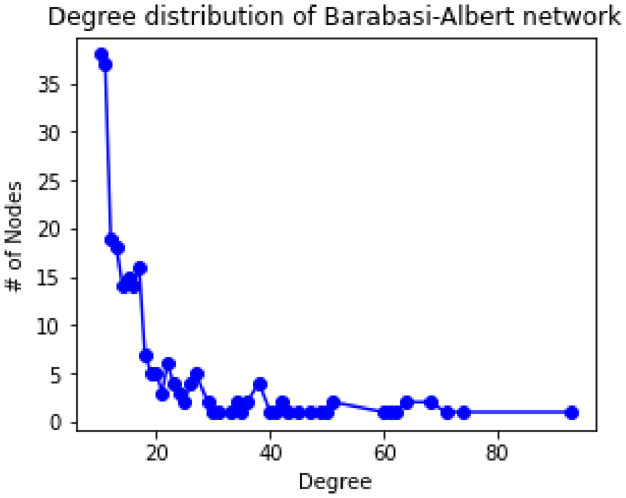


Figure 4: Degree distribution of a Barabasi-Albert network

nodes gets an higher degree than others. Barabasi-Albert graph is known to be a relatively good approximation of social networks nowadays.

3.2 Community detection

The next property we computed is network clustering. Network clustering is an organizing process whose goal is to bring together similar nodes; the result is a partition of the network into a set of communities. The effectiveness of network clustering can be measured by modularity. Modularity is a measure of the structure of networks or graphs that measures the strength of the division of a network into communities (2).
We used the *Louvain's algorithm* as a tool to maximize the modularity of our network. It starts with individual nodes as a community, and for each node in turn, it tries to merge it with another community if this increases modularity. Because it chooses the starting node randomly, one of its strengths is the randomness of this method. This also means that you can get multiple results from the same network partitioning code. To confirm that your partitioning is relevant, it is highly recommended to do a few iterations. If the results are the same, you can certainly say that your partitioning is relevant. This process is called robustness of your method: the more identical your results are, the more robust your method is, and the Louvain algorithm is known as one of the most robust clustering methods. (3)
We chose to do 100 iterations and use the average partition as our main graph.

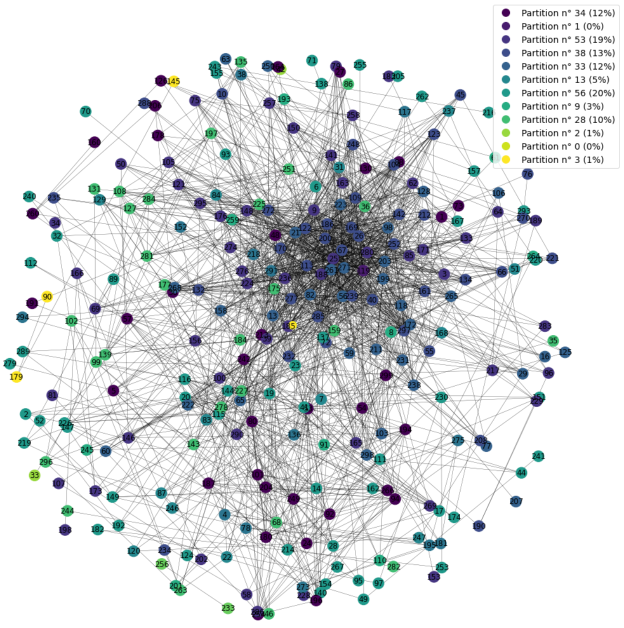


Figure 5: Community detection of Facebook network

Unfortunately, modularity seems to be not really appropriate to any directed graph. To still achieve community detection in our Twitter graph, we need to transform it into undirected network. After obtaining the partitions, we assigned them to the same nodes in the directed network.

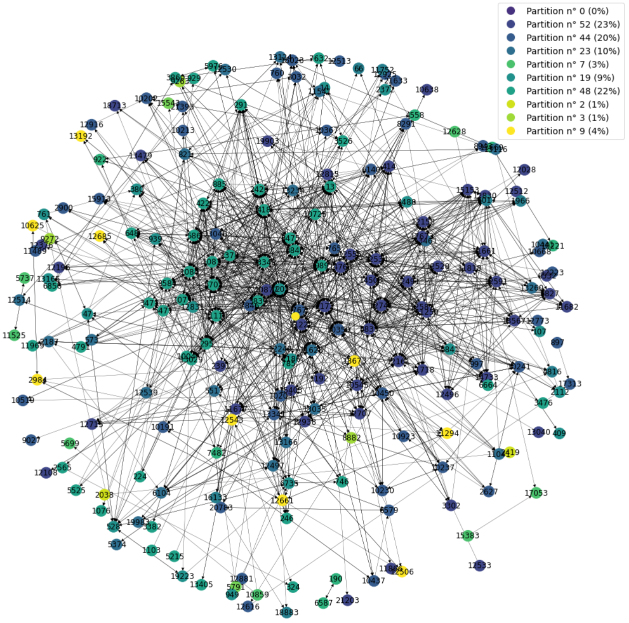


Figure 6: Community detection of Twitter network

3.3 PageRank

The PageRank algorithm is an algorithm initially used by Google to rank pages in their search engine results. It is a way to measuring the importance of website pages. It works by counting the number and quality of links to a page to estimate how important the website is. The rank of a node i is measured by a weight w_i such that $\sum_{i=1}^{N} w_i = 1$. (4)

We will apply this to our directed Twitter Network. The goal is to see if there is a redundancy between the choice of this algorithm and further with propagation algorithms.
For our undirected Facebook Networks, it is a little bit different. Theory seems to say that the PageRank score of a node from an undirected graph tends to the degree of this node, although there is some studies that proves that this is a little bit more complicated. (5)

For the Twitter graph, nodes with the highest rank score are the **2172**, **20** and **10350**. For Facebook, the results give the nodes **25**, **119** and **56**. We will discuss this results later as mentioned before.

4 Start seed set

The start seed set represent the first person aware of the news. We are going to fix the size of the seed set to six percent of the total number of nodes. This represents respectively 17 for the undirected network and 13 for the directed network
Initially, nodes will be randomly selected to start the news spread. Then, the goal being to appreciate a fast propagation, for example a fake news, we are going to define the nodes initially infected to maximize the influence. Those optimal starting nodes are selected by a Greedy algorithm. This one is not the best but is still better than targeting high degree nodes and doesn't require too much computation time.It starts by choosing the best initiator and add it to the seeds nodes. Then it took the second one with the best marginal gain and so on until it has the set of k set nodes which contain six percent of the total nodes. As it optimize each step separately, it only consider the individual propagation of each nodes and not the combination of them. Thus, the optimization is approximate but it require less calculation as it only have to calculate the propagation of $\sum_{i=0}^{k}(n-i) \approx kn$ nodes. The Greedy algorithm gives unfortunately different node choices from the PageRank. However, this is not as surprising, as other studies shows PageRank is only adapted to social networks like Twitter if there is lots of information (6) (7).

5 Diffusion process

We compute the propagation of the news using two diffusion algorithms: the Linear Threshold model and the Independent Cascade model. These two model will simulate, based on the seeds nodes, how would the news spread in the network.

5.1 Independent Cascade Model (ICM)

The independent Cascade Model basically use probability. Each edge have a random weight of $p_{u,v} \in [0,1]$ which is the probability to 'infect' a node. Every time step, the infected nodes can infect his neighbours with the probability $p = 0.1$ associate to each edges starting from an infected nodes. Once a node is infected at time t, it will try to infect its neighbors at time t+1 only then, it can no longer transmit the rumor. The model also assume that someone aware of the news does not forgot it and remains among the 'infected' until the end.

The figures above display the result of different simulation on the two data sets. For both network, we simulate six spread which takes random nodes as initiator and another simulation is starting from the nodes optimizing the influence by Greedy algorithm.
Basically, we can see that the spread behaves in a similar way in both networks. The news spreads quite quickly at the beginning and then decreases as there are less and less people who do not know the information yet. However, we can notice that the number of nodes reached on the undirected network is permanently higher than for the directed one. This is obviously due to the fact that the information is harder to propagate when the edges are directed because it can only be transmitted in one direction and not both.

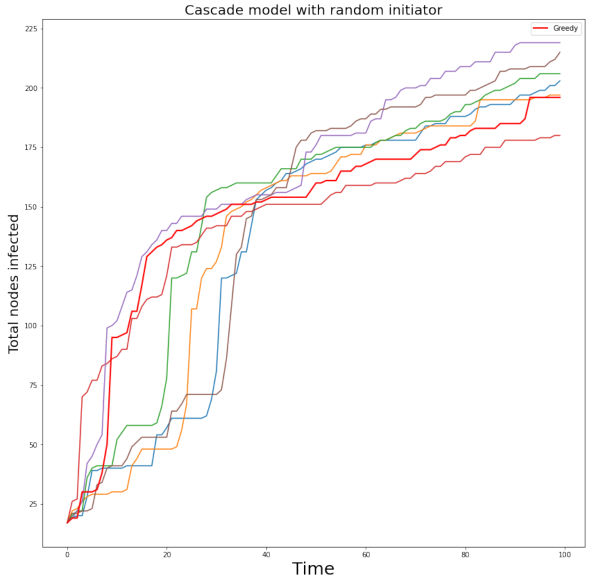


Figure 7: ICM in Facebook Network

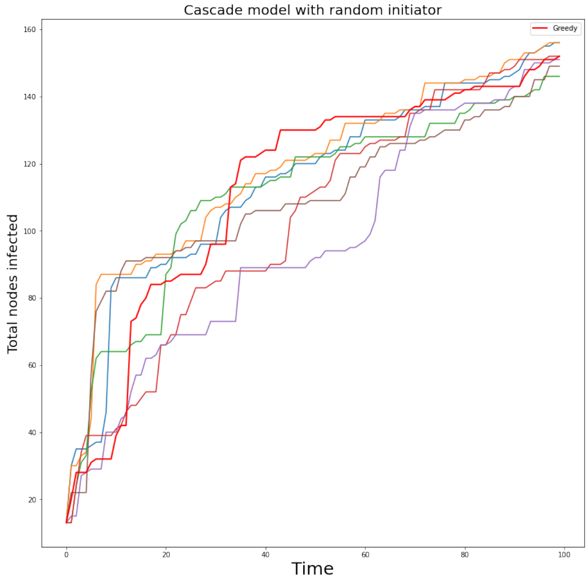


Figure 8: ICM in Twitter Network

Influence maximization does not seem to impact the propagation more than that. The curves corresponding to the propagation from the optimized seeds (in red) are for the two networks located in the average throughout the propagation.

5.2 Linear Threshold Model (LTM)

This model assigns a threshold $\theta \in [0, 1]$ to each node and a weight to each edge starting from an infected node. At each time step, the nodes whose sum of edge weights is higher than the thresholds are activated. Again, it is assumed that an activated node does not forget the news and remains 'infected' until the end.

The figure show the result for the same simulation as for the Independent Cascade model. xxx simulation are made with random initiator and another one start with optimised nodes according to the Greedy algorithm.

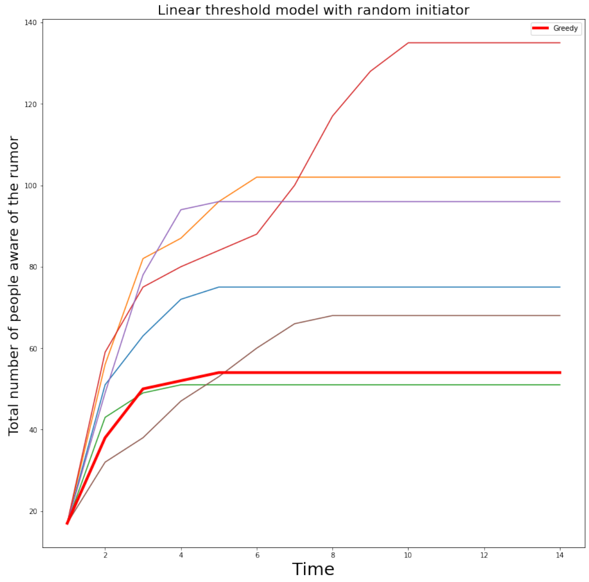


Figure 9: LTM in Facebook Network

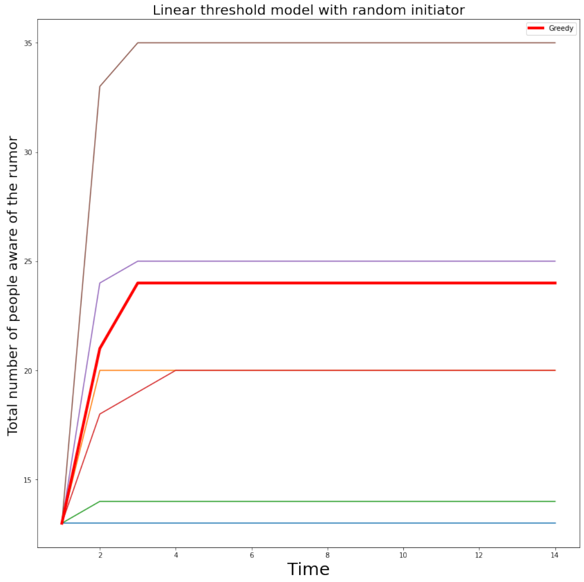


Figure 10: LTM in Twitter Network

The linear threshold model diffusion quickly reaches a plateau for both networks. The news spreads very quickly at first when nobody knows about it and then stops spreading all of a sudden. The model quickly reaches the first nodes before getting stuck in a situation where no more nodes

exceed its threshold and the propagation can no longer evolve. As for the independent model, the spread is much greater on the undirected network as it transmits information more easily.
Once again, the Greedy algorithm does not seem to impact the propagation more than that. The curve corresponding to the spread from the optimized nodes (in red) is in the average.

5.3 Comparison between ICM and LCM

Both models unsurprisingly produce more propagation in an undirected network like Facebook. The linear threshold model is more efficient at the beginning of the transmission but reaches a ceiling very quickly which is low compared to the independent cascade model which reaches more people at the end.

6 Conclusion

These two networks are quite different and by default, the propagation in them is also different. An undirected network should normally reach more people since the information can spread in both directions. The tests carried out confirm this first intuition
However, the difference between Facebook and Twitter does not end there. The results must be qualified in relation to reality. The databases used form networks of 'casual' on twitter and Facebook. Each node has some edges but none really stands out. If the Facebook network looks more or less like this nowadays, Twitter normally has a big disparity in the number of followers per account and our network does not reflect this at all. In addition, these networks do not take into account the comment system either, which can then reach certain nodes without being linked to them. Finally, the activity is not simulated either. There are many more posts on Twitter than on Facebook. All these details to say that while Facebook should reach more people with information from its undirected network, we mostly hear about information posted on twitter these days.

References

[1] Degree distribution
https://www.sci.unich.it/ francesc/teaching/network/distribution.html

[2] https://en.wikipedia.org/wiki/Modularity

[3] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte et Etienne Lefebvre, ≪ Fast unfolding of communities in large networks ≫, Journal of Statistical Mechanics: Theory and Experiment, vol. 2008, no 10, 9 octobre 2008, P10008 (DOI 10.1088/1742-5468/2008/10/P10008, Bibcode 2008JSMTE..10..008B, arXiv 0803.0476)

[4] PageRank, Wikipedia.
https://en.wikipedia.org/wiki/PageRank

[5] GROLMUSZ, Vince. A note on the pagerank of undirected graphs. *Information Processing Letters*, 2015, vol. 115, no 6-8, p. 633-634.

[6] Sigit Priyanta and I Nyoman Prayana Trisna, "Social Network Analysis of Twitter to Identify Issuer of Topic using PageRank" International Journal of Advanced Computer Science and Applications(IJACSA), 10(1), 2019. http://dx.doi.org/10.14569/IJACSA.2019.0100113

[7] https://www.researchgate.net/figure/A-PageRank-algorithm-for-the-social-network-defined-in-Fig-1-Note-that-we-consider-both$_{f}ig3_{2}$21100370

[8] Erika Fille Legara, (2016). A Brief Introduction to Clustering and Community Detection Methods Using NetworkX.
https://notebook.community/eflegara/NetStruc/6.