Marketing information diffusion in an online social network

Victoire Fritsch
MS Data Science
ENSAE ParisTech
victoire.fritsch@ensae-paristech.fr

Stephane Urena
MS Data Science
ENSAE ParisTech
stephane.urena@ensae-paristech.fr

Abstract—The purpose of our project is to understand how to optimize the diffusion of a marketing campaign within a social network. In this paper, we focused on how to chose the most influential users in order to maximize the number of people reached by the marketing campaign at the end of the diffusion. We used a Twitter users network, and focused on the "retweet" relationship between users. We applied the independent cascade model to this empirical network which, according to the literature on the topic, is a good fit for our problem. We identified the initial nodes for the model ("influential nodes") with degree centrality measures, and PageRank measures.

1 Introduction

Wikipedia defines a social network as a "social structure made of individuals (or organizations) called nodes, which are tied (connected) by one or more specific types of inter-dependency, such as friendship, kinship, common interest, financial exchange, dislike, sexual relationship or relationships of beliefs, knowledge or prestige." Unlike physical relationships which are bound by time and space, online social networks increase the proximity between users, and provide communication channels that spread information even quicker. Indeed, people are now able to share information anytime and anywhere through social networks, all of it being inexpensive and reliable. Twitter is a popular microblogging service ("micro" because of the constraint on content size), through which users send and receive text-based posts known as "tweets". Twitter users follow others or are followed. Unlike on most online social networks, following and being followed does not require reciprocation. A user can follow any other user, and the user being followed does not need to follow back.

As part of a marketing strategy, online social networks are thus very effective platforms for companies to spread product information [1]. Indeed many companies have already adopted "seeding strategies" that target influential nodes in social networks, especially to launch new products. Information diffusion is based on trust and knowledge. The users' behavior such as retweeting ('@' followed by a user identifier) or commenting can help spreading information in a social network.

In this paper, we analyze how to ensure the best diffusion of a marketing campaign on a social network, by choosing the right influencing nodes and an appropriate diffusion model. Our analysis is based on Twitter data. Our work was organized in three steps. First, we created the network and analyzed its key metrics. Second, using the network structure and the centrality measures we computed we use the independent cascade model to simulate the diffusion of the marketing information. This step allows us to compare the results of our different simulations and figure out which cen-

trality measure was the most appropriate. Third and final, based on the best centrality measure(s) we list the top-rank nodes that we will consider are the most influential nodes in our Twitter network.

This paper is organized as follows. We further motivate our study with a review of literature on the topic in Section 2. Our experimental design and methodology is described in Section 3. Section 4 is dedicated to reviewing our results and findings. Finally in Section 5 we conclude.

2 Review of literature on the topic

In this part, we review (i) the literature on diffusion models in online social networks, and (ii) literature on identification of influential nodes in such networks, in order to determine which ones we will use in our experiments.

2.1 Discussion on diffusion models

Social networks and their properties as a way to represent the relations and communications within and between different social groups, and information propagation within them, have been an object of active studies. Milgram [2] shows that the average path length between two Americans is 6 hops, and Pool and Kochen [3] provide an analysis of the small-world effect.

As online social networks are gaining popularity, sociologists and computer scientists are beginning to investigate their properties. Moreover, online networks are focused on sharing information, and as such have been studied extensively in the context of information diffusion. More specifically, a key characteristic of diffusion models is the correlation between the number of friends engaging in a specific behavior, and the probability of adopting the behavior.

According to Li et al. [4]; the literature related to these issues can be classified into two categories: explanatory diffusion models and predictive diffusion models. Explanatory models answer the question "Why" on a posteriori data. The objective is to understand why information has propagated this way, what were the interactions, etc. On the other hand, Predictive models refer to the question "Where", related to the future diffusion of information. Li et al. take the example of a user A who has two friends, B and C, in a social network. Both B and C are influential users. If A posts some information, both B and C will each have a different perspective on the information, which will influence how they respond and whether they further propagate it through the network. These factors help understanding the prediction of information diffusion. The information diffusion models based on Li et al.'s view are depicted in Figure 1.

Our project focuses on how influential users play a role in the diffusion of a marketing campaign on an online social network. Hence, predictive models are more appropriate for our experiments. Two of these models are seminal models: the Independent Cascade model (ICM) [5] and the Linear Threshold model (LTM) [6]. These two models are based on a directed graph where each node can be activated or not.

These two diffusion models (LTM and ICM) were used and compared with online social networks by Samir Akrouf et al. [7], based on an initial set of active nodes that had been previously selected. They noticed that choosing the diffusion model is relevant to the network and its type of relationships. LTM model is better on the egocentric explicit network (created based on explicit relationships from Flickr service) to get better influence predictions, while ICM performs better on the implicit network (created by commenting relationship from YouTube service) with stronger ties, since it is based on the interactions between nodes. Thus, the ICM diffusion model fits better the real-world diffusion process in a retweeting network. Therefore, we chose to use the ICM to perform our simulations.

2.2 Discussion on node selection

According to the literature on the topic [8], [9], [10], the ability of influencing users to initiate a large-scale spreading is mostly due to their privileged locations in the underlying social networks.

The most straightforward measurement of influence is using centrality-based measurements. In recent years, an increasing number of methods have been adopted to ranking node's influence in a social network, among which degree centrality [11], betweenness centrality [12] and PageRank [13].

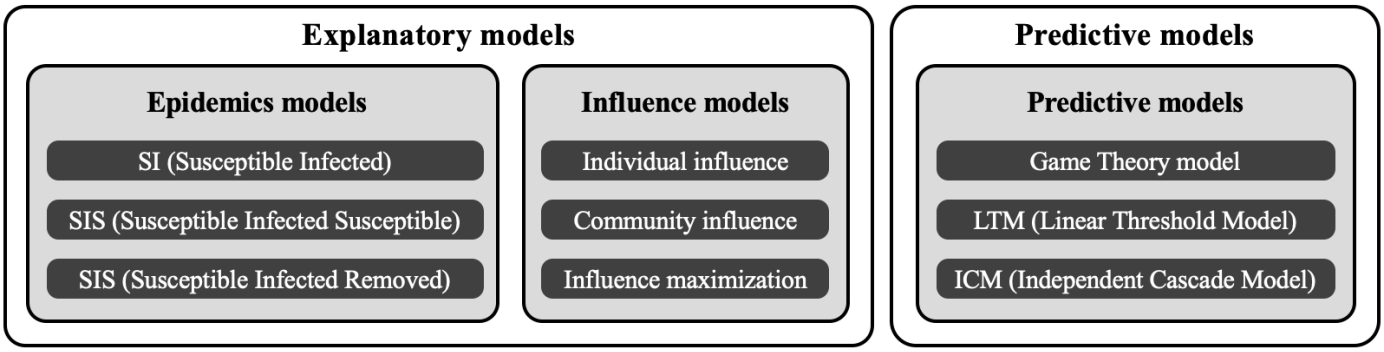

Figure 1: Categorization of Information Diffusion Models by Li et al.

Mislove et al. [14] have observed a high degree of reciprocity in directed user links, leading to a strong correlation between user in-degree and out-degree. They found that online social networks contain a large, strongly connected core of high-degree nodes, surrounded by many small clusters of low-degree nodes. This suggests that high-degree nodes in the core are critical for the connectivity and the flow of information in these networks. Li et al. [15] proposed a descriptive diffusion model to take dependencies among the topics into account and identify the most influential nodes for specific contagion. They applied the proposed model both on simulated data and on a Twitter dataset (ISIS) to predict diffusion. Kwak et al. [10] ranked users by (i) number of followers, (ii) PageRank and (iii) betweenness centrality to identify influential nodes on Twitter. Akrouf et al. [7] selected the set of active nodes for their diffusion models based on the structural metrics of nodes, where nodes with high overall degree, out degree and between centrality values were chosen.

As the models just cited show, centrality-based measurements are also widely used to evaluate a node's influence. In our project, we thus chose to compare degree-centrality (in-degree and out-degree measures), betweenness-centrality as well as PageRank to identify influential nodes.

3 Experimental design and methodology

The methodology for our experiments unfolds in 3 steps:

1) Construction and visualization of the network based on the Stanford SNAP Twitter data-set, and computation of the networks' metrics (centrality, PageRank).
2) Application of the ICM to the aforementioned network, with different sets of initial nodes determined based on the network metrics, and a set of random initial nodes.
3) Ranking of the results by number of active nodes at the end of the propagation in order to determine the most influential nodes.

3.1 Network visualization and analysis

3.1.1 Data resource

We chose to build our network from an already existing one, based on the well known social network Twitter. The data-set was extracted from the Stanford SNAP library. Called the "Higgs data-set", it has been built after monitoring the spreading processes on Twitter before, during and after the announcement of the discovery of a new particle with the features of the elusive Higgs boson on 4th July 2012. The messages posted in Twitter about this discovery between 1st and 7th July 2012 are taken into account.

The data consists of four directional networks that have been extracted from user activities in Twitter as:

— re-tweeting (retweet network)
— replying (reply network) to existing tweets
— mentioning (mention network) other users
— friends/followers social relationships among user involved in the above activities
— information about activity on Twitter during the discovery of Higgs boson

This extracted diffusion network contains 456,626 vertices (users) and 14,855,842 edges (relationships) between them. The relationship between vertices include retweets (RT), replies (RE) and mentions (MT). Thus, the data-set is directed.

On Twitter, a user can follow any other user, and the user being followed need not follow back. Thus, as stated by Kwak et al. [10], it is the retweet mechanism that empowers users to spread information of their choice beyond the reach of the original tweet's followers. For this reason, we decided to focus on the retweet relationship between users, which represents 256,491 vertices and 328,132 edges.

3.1.2 Visualizing the network

After building the Twitter diffusion network, the network structure was analyzed both visually and quantitatively. The network was visualized using $GePhi$ [1], an open-source platform for visualizing and manipulating large graphs. The platform provides several ways of spatializing data, depending on their nature and their size. In our project we chose to use the $ForceAtlas2$, a force-directed layout that aims at transforming the network into a map by integrating different techniques such as the Barnes Hut simulation, degree-dependent repulsive force, and local and global adaptive temperatures.

The metrics that we considered relevant for our experiments following our literature review were the in-degree, the out-degree, the overall-degree, the betweenness centrality and PageRank.

The in-degree, the out-degree and the overall-degree are used to determine degree centrality. In-degree and out-degree are specific for directed graphs (where edges have directions). The in-degree represents the number of edges incoming to a vertex (in our specific case : how many people retweeted this vertex). The out-degree: the number of edges outgoing from a vertex (in our case, how many people were retweeted by this vertex). Betweenness centrality quantifies the number of times a node acts as a bridge along the shortest path between two other nodes. Finally, PageRank works by counting the number and quality of links to a node to determine a rough estimate of the importance of the node. In short PageRank is a "vote", by all the other nodes, about how important a node is. An edge to a node counts as a vote of support. If there's no link there's no support.

These metrics were computed using the $Networkx$ Python package, designed for the creation, manipulation, and study of the structure, dynamics, and functions of networks. All of the nodes in the data-set were ranked based on these measures, and the top-ranked nodes for each measure were selected as the initial set of active nodes. These initial active nodes are used to initiate the information diffusion with the Independent Cascade Model.

Betweenness centrality is a slow calculation. The algorithm used by $Networkx$ is $O(VE)$ where V is the number of vertices and E the number of edges. In our situation, this represents a calculation in $O(84,162,904,812)$. We tried different means to parallelize the computation, but gave up because of time constraints. We chose not to apply it to a small sample of our data because our network is quite concentrated and in the case were the important nodes are not in our sample, the message could be completely different. For these reasons, results on betweenness centrality are not displayed in the rest of the paper.

3.2 Independent Cascade Model application

The model, first introduced by Kempe et al. in 2003 [5], starts with an initial set of active nodes A_0, and diffusion unfolds according to the following randomized rule:

- When vertex v becomes active in step t, it is given a chance to activate each of its inactive neighbors with a probability $p(v, w)$ for the activation of vertex w.
- If v succeeds to activate its neighbor w, then w will become active in step $t + 1$, and will be added to A_s to form the new active vertices set A_{s+1}.
- Then, w will adopt the same activation action to activate its inactive neighbors.
- Whether or not v succeeds in activating w in round t, it cannot make any further attempts to activate w in subsequent rounds.

1. Gephi is developed in Java and uses OpenGL for its visualization engine.

— The process runs until no more activations are possible.

Previous research have improved the ICM by reducing the computation cost of the algorithm [16] and the accuracy of the probability applied to the model [17]. However in our project we decided to consider only the basic ICM. We also set the probability of succeeding to activate a vertice to 0.5. This an adoption rate that is quite high, but we assume that our marketing campaign will be really attractive.

Selecting an adequate set of initial adopters is key for our diffusion model. As previously mentioned, the set of initial nodes is selected based on degree centrality. PageRank is also considered as an additional measure since it is commonly used in the literature to estimate the importance of a node. Five experiments were conducted, each one using one of the metrics described in part 3.1.2. We used for this experiment the *NDlib*[2] package (Network Diffusion Library) on Python. Built upon the *NetworkX* python library, NDlib allows simple and flexible simulations of networks diffusion processes.

Finally, we compared the results of these experiments, in order to determine which were the most influential nodes, meaning the ones that would maximize the diffusion of the information (i.e. the final number of active nodes).

4 Results

4.1 Network visualization and analysis

As previously mentioned, the imported Twitter Higgs retweet network contains 256,491 vertices and 328,132 directed edges. Both Figure 2 and Figure 3 represent the network graph, showing the connections between users and the key clusters. A vertex is created when a user posted an original tweet. An edge is created when a user retweeted the original message.

Modularity is a measure of the structure of networks or graphs, that was designed to measure the strength of division of a network into modules (groups, clusters or communities). In our situation, modularity was used for detecting community structure in our network. From Figure 2 we can clearly distinguish 6 different communities in our Twitter Higgs diffusion network.

The eccentricity is a node centrality index. The eccentricity of a node v is calculated by computing the shortest path between the node v and all other nodes in the graph, then the "longest" shortest path is chosen, and its reciprocal is calculated. Thus, an eccentricity with higher value assumes a positive meaning in terms of node proximity. From Figure 3, we can see that our Twitter network is quite concentrated. The pink nodes have an eccentricity of 1 (they only have one connection), and the two shades of green represent eccentricities of 14 and 15. These two shades of green represent 17% of the nodes. When adding all the results above 2, we reach 31.5% of the nodes. So these 31.5% of the nodes concentrate all the connections.

Measure	Value
Number of vertices	256,491
Number of edges	328,132
Graph density	4.988e-06
Diameter	23
Modularity	0.796
Distance	7.937
Number of strongly connected components	255,002
Number of weakly connected components	13,199
Average overall degree	2.5527
Average out-degree	1.2793
Average in-degree	1.2793

Table 1: Twitter Higgs diffusion network basic metrics

Then, the network's basic metrics were computed with *Networkx*, as displayed in Table 1. The objective of these metrics was to characterize our network. In Table 2, we ranked the nodes by overall-degree centrality, out-degree and in-degree centrality, and PageRank and displayed for each measure the top ten vertices. All user names were anonymized and replaced by numbers. Even though overall-degree and in-degree are similar in the top 10 nodes, there are a few differences in the top 20 so we kept both measures for our experiments.

2. *NDlib* is a result of two European H2020 projects: *CIMPLEX* "Bringing CItizens, Models and Data together in Participatory, Interactive SociaL EXploratories": under the funding scheme "FETPROACT-1-2014: Global Systems Science (GSS)", grant agreement 641191 and *SoBigData* "Social Mining Big Data Ecosystem": under the scheme "INFRAIA-1-2014-2015: Research Infrastructures", grant agreement 654024.

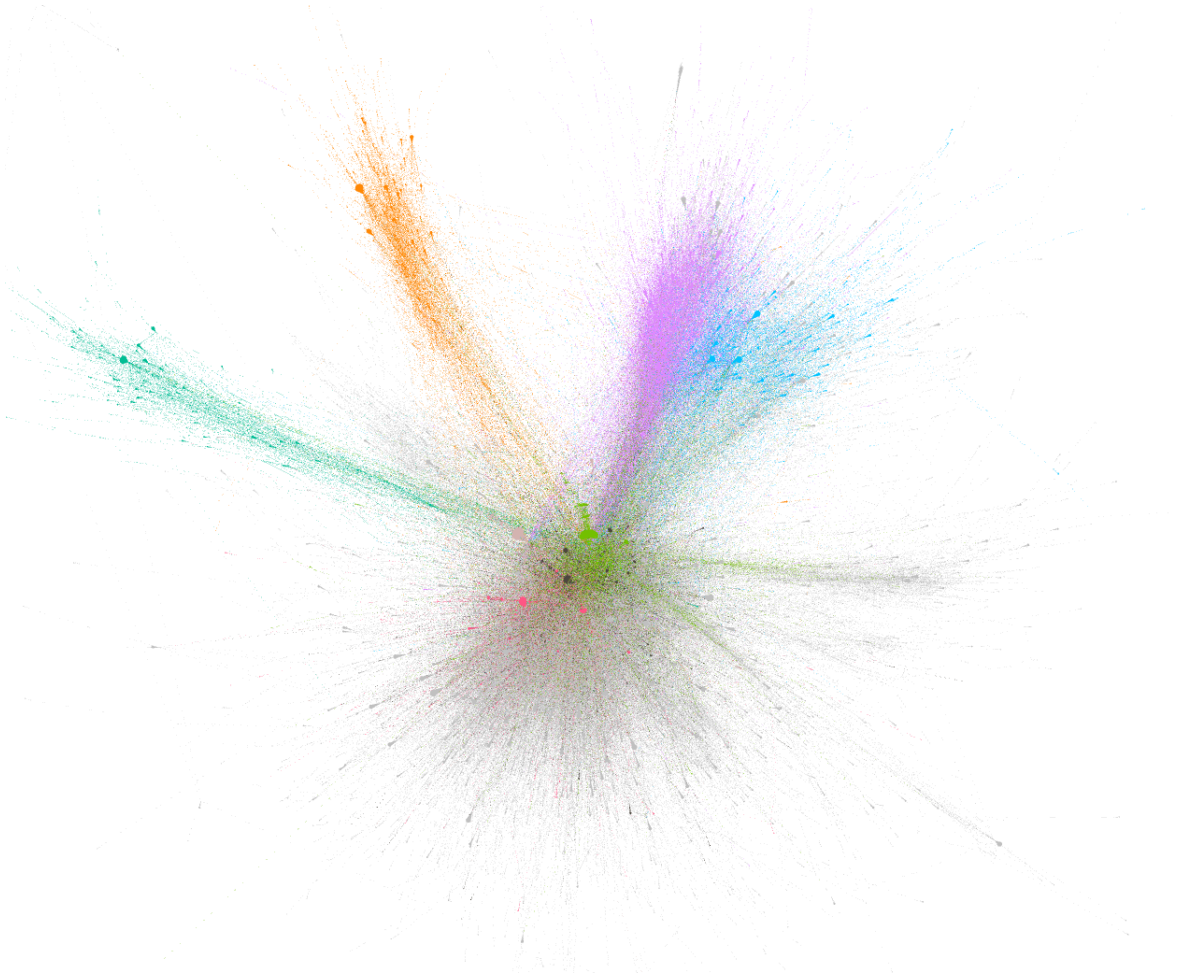

Figure 2: Network visualization partitioned by Modularity

Rank	Overall-degree	In-degree	Out-degree	PageRank
1	88	88	38,535	88
2	677	677	181,190	2,342
3	1,988	1,988	81,405	64,911
4	349	349	54,301	39,420
5	3,571	3,571	64,911	1,988
6	9,964	9,964	27,705	677
7	5,226	5,226	53,508	3,998
8	519	519	232,850	134,095
9	2,567	2,567	492	169,287
10	19,913	19,913	52,204	2,567

Table 2: Top 10 nodes based on the different measures

4.2 Experiments results

The ICM was conducted on the previously built network for each of the five measures discussed in section 3.2 and according to the protocol defined in the aforementioned section. In addition, five sets of initial active nodes were tested for each measure : the top 10, top 20, top 50, top 70 and top 100 nodes.

For each metric and each of the five sets of initial active nodes the ICM was run 200 times. The diffusion of the information was quantified according to the average number of active nodes at the end of the process for each of the four sets and five metrics. This is a was to quantify the ”influence” of the set chosen to initiate the process. A set of randomly selected nodes was also used to

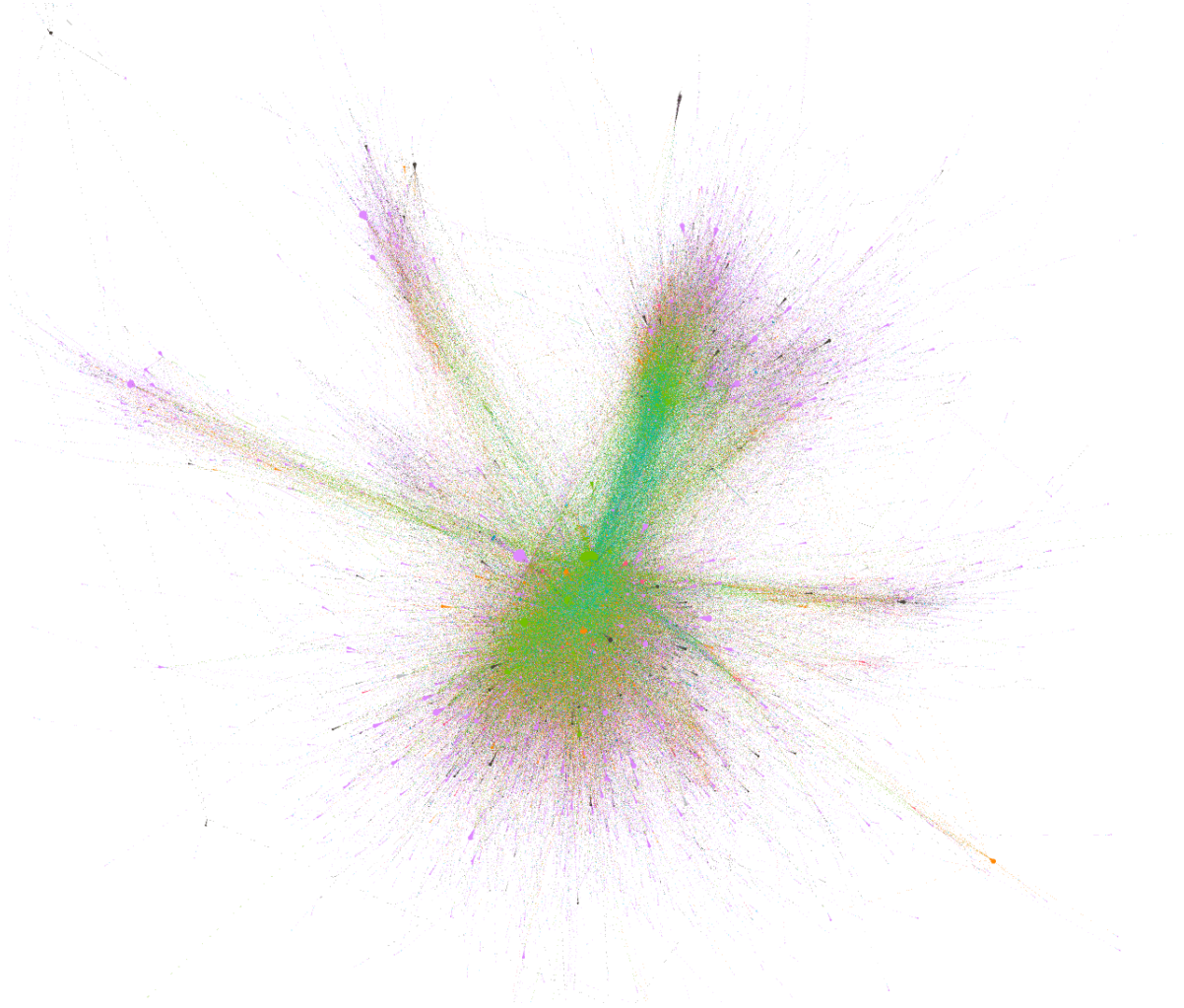

Figure 3: Network visualization partitioned by Eccentricity

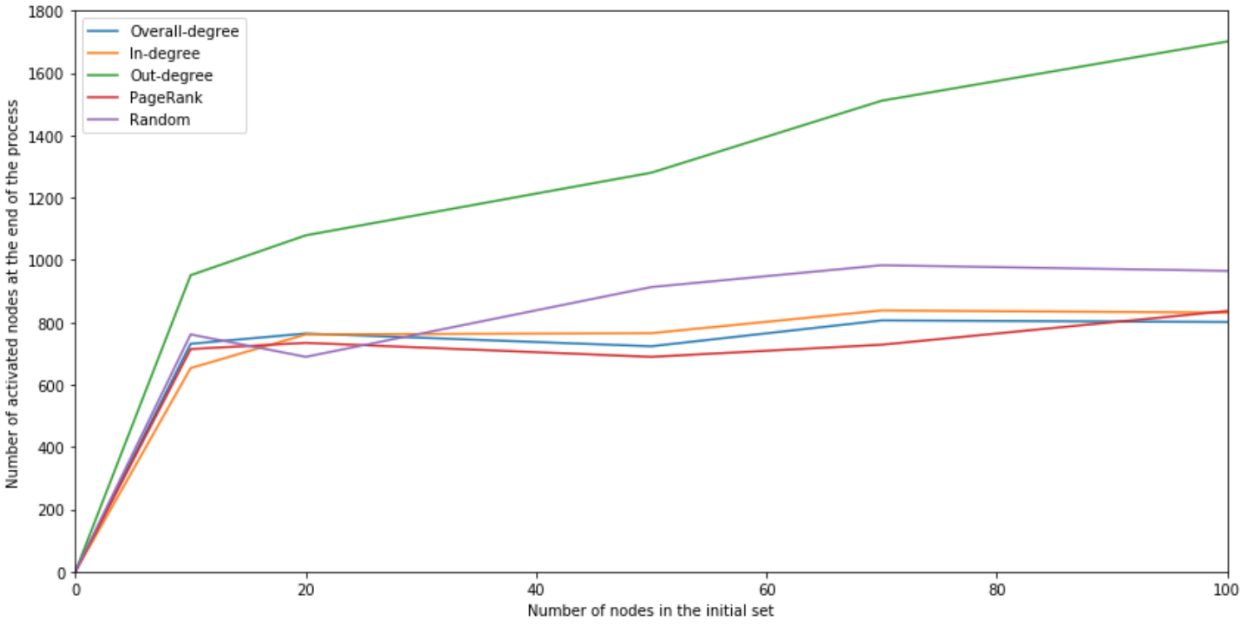

Figure 4: Results of the ICM simulations

compare to our results.

The results of these 35 experiments are presented in Figure 4. It can be seen that the results converge to approximately the same number of activated nodes (more or less 800) when the initial set of active nodes is chosen based on overall-degree, in-degree, and PageRank As the convergence numbers are already reached with an initial set of 20 nodes, it is not necessary to input more. Indeed, the total number of activated nodes includes the ones from the initial set. It means that the number of nodes activated in the process is actually decreasing with each bigger set of active nodes.

The randomly selected nodes give a slightly better performance, with a convergence around 900 activated nodes. The measure that gave by far the best performance is the out-degree, with nearly 1,800 nodes activated by only 100 initial nodes.

4.3 Ranking of the results

In our project, we determine if a set of nodes is influential or not based on the number of activated nodes at the end of the process. Based on the results of Figure 4, the best measure for seeding was the out-degree centrality. We performed additional tests using this measure as our seeding strategy, running simulations with the top 1,000, top 5,000, top 10,000 and top 20,000 nodes.

The results are displayed in Figure 5. We showed both the total number of activated nodes at the end of the process, and the net number of activated nodes during the process (computed as the total number of activated nodes at the end minus the number of nodes in the initial set). We observe that the net number of activated nodes increases fast at the beginning, up until 1,000 nodes. Then it continues to increase but at a slower rate.

We can imagine that at some point this net number of activated nodes will converge. However, in our situation it is not relevant to have a further look. Indeed, if we contact more than 10% of the targeted population, we will not be in the same marketing perspective.

5 Conclusion

To conclude, in our project we compared different seeding strategies to optimize the diffusion of a marketing campaign on a social network by using Twitter data. The data we used was users' retweets regarding the announcement of the discovery of the Higgs boson in June 2012. We simulated diffusion on this network with the help of the Independent Cascades Model, and used centrality measures and PageRank to determine our initial sets of nodes. To determine whether or not a set of nodes was "influential", we looked at the final number of nodes it activated.

On social networks, users influence each others through social interactions. Thus, a proper seeding strategy could increase the number of customers reached by a marketing campaign, by reaching in this way niche markets.

In their study of marketing product purchases, Leskovec et al. [18] concluded that the probability of buying a product increases with the number of recommendations received (thus the number of social interactions), but at some point (reached quickly) it decreases and converges to a low probability. But as underlined by Zhang et al. [19] retweeting is not recommending, it does not require the same effort. As a "low-cost" behavior, retweeting could effectively promote the information.

But to conclude fairly we need to mention the numerous limitations of our study. First, the diffusion model is basic and simple. The probability of converting attributed to each node is constant, and each node reached but not infected is removed. This is not really what happens in real world. Second, the network that we chose (the Higgs Twitter retweet network) was not completely adapted to our needs. Even though its structure matched perfectly our expectations (it represents the diffusion of an information on a social network), the topic is far away from our considerations. People interested in sciences have their own network structure, while users sensitive to marketing information probably have a very different network structure.

Thus, to improve our project the first thing would be to obtain more appropriate data. The actual diffusion network of a marketing campaign

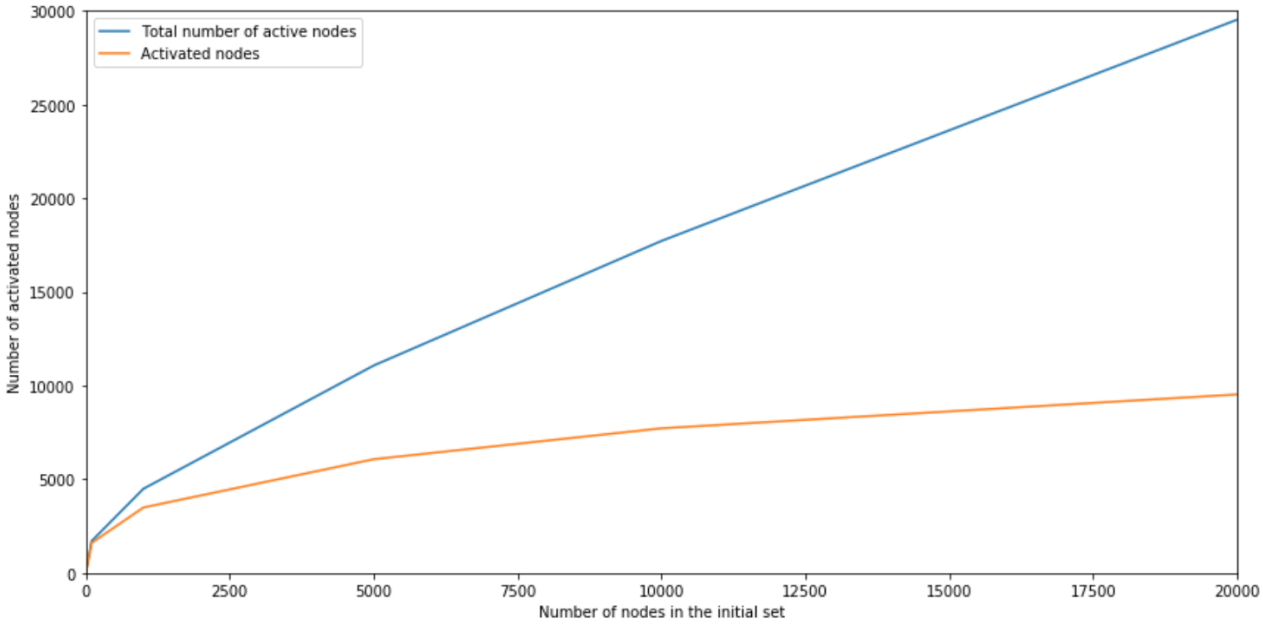

Figure 5: Results of the ICM simulations - focus on out-degree

or a product launch would suit better the topic. Moreover, the parameters of the simulations could be fine-tuned. For example instead of having a constant probability for each node to turn active, we could adapt the probability to each node according to its attributes. Finally, a non-anonymized data-set would allow some analysis on the nodes' background and real-life importance.

6 Contributions

Mutual work:
— choice of topic and dataset
— discussion on models
— network visualization

Stéphane:
— network metrics
— specific focus on betweenness, which did not give the expected results because of time constraints but was a consequent topic in our project

Victoire:
— literature review
— ICM simulations

References

[1] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *The eighth ACM SIGKDD international conference*, pages 61–70. ACM, 2002.

[2] S. Milgram. The small world problem. In *Psychology Today*, 2(60), 1967.

[3] I. Pool and M. Kochen. Contacts and influence. In *Social Networks*, pages 1-–48, 1978.

[4] M. Li, X. Wang, K. Gao and S. Zhang. A survey on information diffusion in online social networks. In *Models and methods*, Information 2017, 8, pages 1—21.

[5] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 137-–146, ser. KDD '03, 2003.

[6] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *Presented at the 2010 IEEE 10th international conference on data mining (ICDM)*, pages 599—608, IEEE, 2010.

[7] S. Akrouf, L. Meriem, B. Yahia and M. N. Eddine. Social network analysis and information propagation: A case study using Flickr and YouTube networks. In *International Journal of Future Computer and Communication*, pages 246—252, 2013.

[8] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. In *Journal of Consumer Research*, pages 441—458, 2007.

[9] H. Yoganarasimhan. Impact of social network structure on content propagation: A study using YouTube data. In *Quantitative Marketing and Economics*, pages 111—150, 2011.

[10] H. Kwak, C. Lee, H. Park and S. Moon. What is Twitter, a social network or a news media? In *Presented at the 19th international conference on World wide web*, page 591, ACM Press, 2010.

[11] R. Albert, H. Jeong and A. L. Barabási. Internet: Diameter of the world-wide web. In *Nature*, pages 130—131, 1999.

[12] F. C. Freeman. Centrality in social networks conceptual clarification. In *Social Networks*, pages 215—239, 1978.

[13] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Com- puter Networks and ISDN Systems*, pages 107—117, 1998.

[14] A. Mislove , M. Marcon , K. P. Gummadi , P. Druschel and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, 2007.

[15] Q. Li, B. Kailkhura, J. J. Thiagarajan, Z. Zhang and P. K. Varshney. Influential node detection in implicit social networks using multitask Gaussian copula models. In *Presented at the NIPS 2016 time series workshop*, pages 27–37, 2017.

[16] M. Kimura, K. Saito, R. Nakano and H. Motoda. Extracting influential nodes on a social network for information diffusion. In *Data Mining and Knowledge Discovery*, pages 70–97, 2009.

[17] M. Kimura, K. Saito, R. Nakano and H. Motoda. Finding influential nodes in a social network from information diffusion data. In *Social computing and behavioral modeling*, pages 1–8, 2009.

[18] J. Leskovec, L. A. Adamic and B. A. Huberman. The dynamics of viral marketing. In *Transactions on the Web*, pages 1—39, ACM, 2007.

[19] L. Zhang, M. Luo and R. Boncella. Product information diffusion in a social network. In *Electronic Commerce Research* , Aug. 2018.