STOCHASTIC INFLUENCE MAXIMISATION PROBLEM

Aelenei Vlad Stefan
ICT4SS
Politecnico di Torino
s286504@studenti.polito.it

Balan Rares Alexandru
ICT4SS
Politecnico di Torino
s290175@studenti.polito.it

Davide Brusco
ICT4SS
Politecnico di Torino
s292470@studenti.polito.it

Valletta Federico
ICT4SS
Politecnico di Torino
s287359@studenti.polito.it

June 18, 2021

ABSTRACT

The aim of this study is to find the best-limited set of nodes, seed, such that the number of finally influenced social network individuals is maximized. Influencers spread information to all the directly connected individuals, from them a cascade propagation starts, as consequence each active node forwards information to its neighbours. The influence process can be of different types, in Stochastic Influence Maximisation Problem (S-IMP) a threshold-based propagation model is used, in which a higher influence increases the probability the receiving individual is activated.
The evolution of the information spreading is performed across one time period. This choice is due to some assumptions and simplifications that are adopted in the development of the project.
The methods adopted to find better solutions are a linear programming approach and a heuristic one. The basic result is that the heuristic solution is obtained in a shorter time than the linear programming approach, it is not optimal but anyway feasible.

1 Introduction

Nowadays social networks have an important role in our community because of their easy availability in everyday life thanks to the diffusion of smartphones and tablets. It is mostly young people who spend more time a day using this technology and social media, but some networks are very popular even among the older generation. In fact, out of a population of more than 7 billion people, mobile devices are accessible to more than 5 billion people, 67% of the earth population. Instead, people accessing the internet are more than 4 and a half billion, 60% of the world population. From this data, it is retrieve that active users on social networks are 4.14 billion, an increase of over 10% compared to last year. Also in Italy this technology is very used: about 58% of Italians use social networks, the average time spent on social media is 2 hours a day and 98% of users access social networks from smartphones [1].
The results of this project can be exploited to plan an advertising campaign. Social networks are a good and fast way to reach a large number of individuals starting from a relatively small number of influencers. This environment allows to connect anyone who has a common interest, this process gives a strong contribution to the spread of commercials and consequently a success of the campaign. These kinds of activities are advertisements that appear on platforms such as Facebook, Instagram, YouTube, Pinterest, Twitter or other media channels that exist in the digital space. Nowadays every company must consider social networks as a primary tool for advertising and exploit them thanks to their similarities but especially for their differences, to feed even more the catchment area. Social networks have proved essential to promote a new product or service, increase brand awareness, customer loyalty, collect feedback and content from users, generate income, increase user involvement, advertise a future event and increase sales of a product that is already on the market. According to the Social Media Examiner research [2], 97% of small businesses use social media to attract new customers.
The Influence Maximisation Problem (IMP) is generally NP-Hard also for relatively small networks, so a linear

programming approach is not always possible due to the waste of time and computational efforts, in favour of a heuristic approach. In Stochastic Influence Maximisation Problem (S-IMP), the problem complexity is lower due to some assumptions and hypotheses, anyway, both methods are used and compared.

2 Similar models in literature

The S-IMP is inspired by the work of E. Güney [3] about the problem of the identification of a set of key individuals that maximizes the node influence coverage in a social network. In this paper, the IMP is treated as a stochastic variant of the maximal covering location with an independent cascade influence model. The authors exploit a branch and cut algorithm based on the Benders decomposition and on its relation with submodular cuts to define network subsets.
A different approach to the problem is treated in the A.Cuzzocrea's paper [4] in which a deep learning method is used to perform the Social Influence Analysis (SIA), the phenomenon that describes the spreading of opinion across the population. Considering a user u and its neighbours v, the state of u at the end of a time interval is predicted, given the statuses of its surrounding neighbours.
An interesting point of view is shown by D. Kempe's paper [5] in which the goal is to try to convince a subset of individuals to adopt a new product or innovation and to find which is the target set. The problem of finding the best set of influential nodes is NP-Hard, for this reason, it is necessary to develop an approximation to provide an efficient resolutive algorithm. In 63% of the cases, the greedy algorithm has a solution comparable with the deterministic one and with a significant save of time. Computational experiments on large collaboration networks show that the proposed algorithm significantly outperforms the node-selection heuristic based on the well-studied notions of degree centrality and distance centrality from the field of social networks.

3 Problem statement and definition of the model

3.1 Variables, parameter and objective function

In our S-IMP, the social network is represented as a graph G with N nodes. The seed is the set of K nodes, used to start the advertisement campaign and belonging to the ensemble V, while x_i, with $i = 0, 1, 2...K$, is a general node influenced by the seed. Referring to the E. Güney's paper exposed in Section 2, the defined objective function (OF) is

$$max \left\{ \sum_{\omega \in} p_\omega \cdot \sum_{i \in V} x_i^\omega \right\}$$

which maximizes the expected number of influenced nodes by a specific seed. To achieve this purpose is necessary to investigate the sum of the multiplications between the scenario probability p and the sum of influenced nodes x_i. This operation is performed for each scenario ω which belongs to the ensemble .

3.2 Graph structure

The project works with an oriented and weighted graph that represents the analyzed social network. The more realistic data structure representing a social network is a scale-free graph, where users are connected according to the Preferential Attachment (PA) rule proposed by Barabasi-Albert [6]. In this model, individuals are more likely to connect themselves to a hop with a large degree. The node degree follows a Heavy-Tailed distribution in which a few nodes are hubs and are highly connected with a low clustering coefficient, from this last characteristic is derived general random graph structure. The most interesting property is the small world effect: even for a large network, each hop is easily reachable. During the development of this research, more typologies of graphs are explored, all of them have some common features. Each node is an individual that takes place in the social network, where for each of them is defined a weight threshold whose overcoming allows their influence. Those values are generated starting from the order of the outcoming arcs from the node. To it, a random noise value, in the range including the $\pm 20\%$ of the node degree, is added, it is useful to include randomity in the graph building process. In the used graph a node is defined as activated if it is chosen to start the influence process, while if it joins in it, it is defined as influenced. In addition to that for each arc, there is also a connection in the opposite direction to show that the influence can be in both directions, from a theoretical point of view. The connections between nodes are generated randomly during the graph generation. The three typologies of data structures used are:

- An easier graph: shown in Figure 1, it is built ad hoc to approach the problem in a simple way and to understand its basic behaviours. It is composed of 14 nodes, two of them are connected from each other and they have 6 others arcs to as many nodes. From this structure, some easy prediction can be done about seed composition.

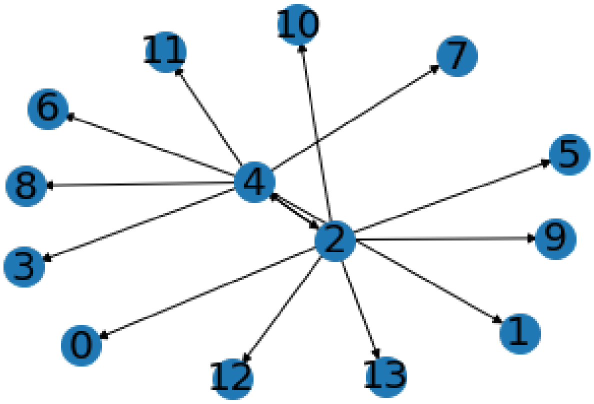


Figure 1: The easier graph used in the project.

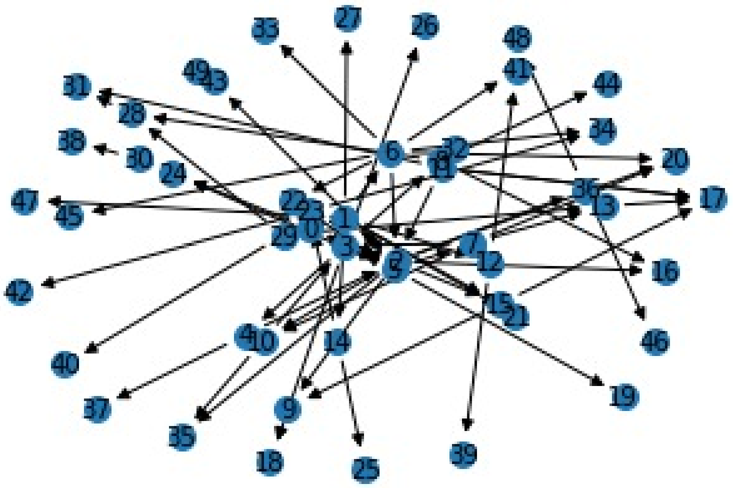


Figure 2: Scale free topology graph, N = 500.

- A second graph is a scale-free topology network: its degree distribution follows a negative exponential, which means that there are a few nodes with a large value of connection and vice versa. An example, shown in Figure 2, is made up of 500 nodes and it is used to obtain more realistic results that can be easily managed. It is built using the NetworkX python library and then adding the weights.

- The third graph is the most complex graph retrieved by The Konect Project website [7], in this document, for simplicity, it is called konect. It is a directed and unipartite graph, which means that all the nodes are of the same nature. In the data structure, the weights are not provided but they are added during the project development. It does not contain directed cycles and loops and it contains reciprocal edges. It is made up of 874 hops, the maximum node degree is 118 while the mean value is 4.240. The clustering coefficient is c is 0.192, which means that the average edge density around a node is low. In Figure 3 are reported the graph topology and the degree node distribution.

Given an oriented graph $G = (V, E)$, where V are the set of nodes and E are the set of arcs, the predecessors $\hat{v}$ of $v \in V$ are all the hops such that the link $e \in E$ starts from $\hat{v}$ and arrives to v.
Given an oriented graph $G = (V, E)$, where V are the set of nodes and E are the set of arcs, the successors $\hat{v}$ of $v \in V$ are all the hops such that the link $e \in E$ starts from v and arrives to $\hat{v}$. Referring to Figure 1: considering, for instance,

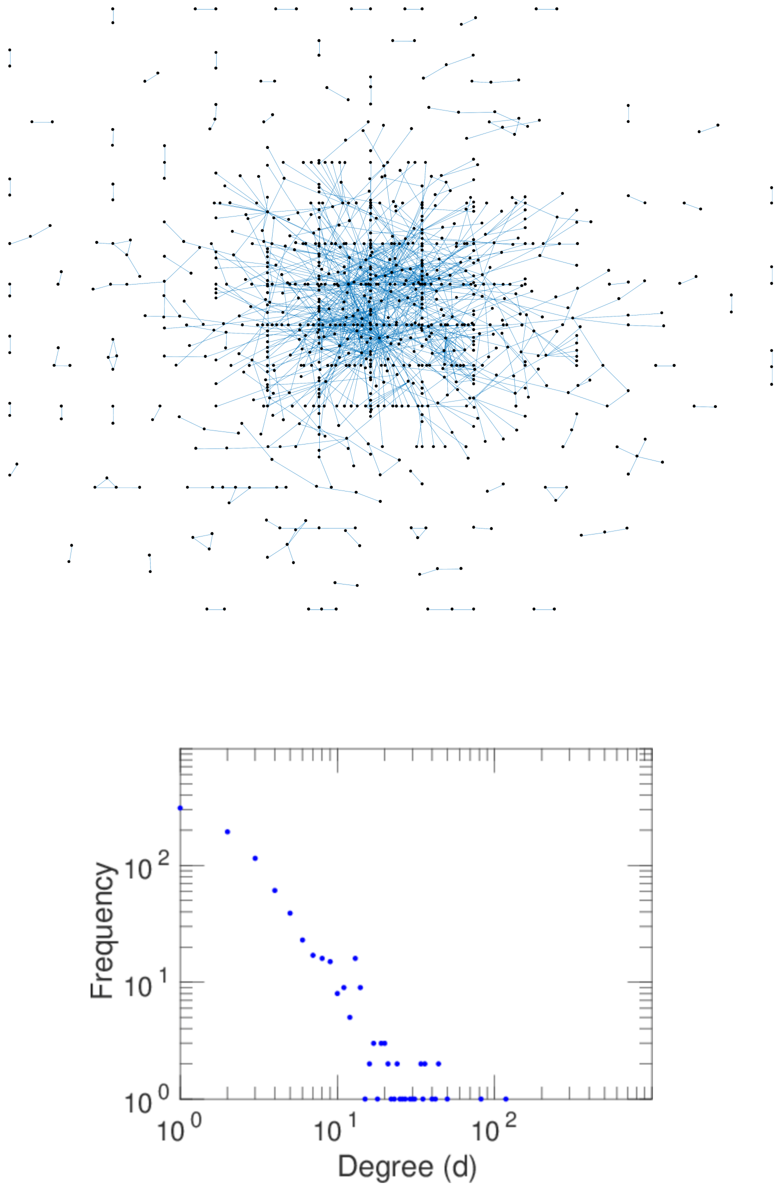


Figure 3: Above konect topology and below node degree distribution.

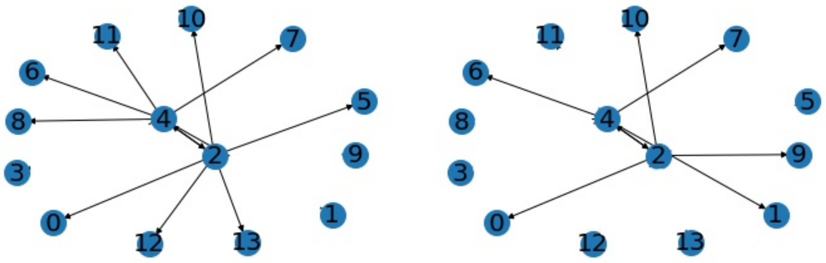
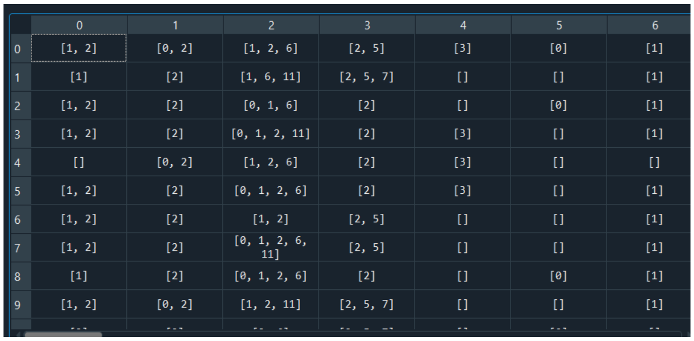


Figure 4: Examples of two scenarios of a simple graph.



	0	1	2	3	4	5	6
0	[1, 2]	[0, 2]	[1, 2, 6]	[2, 5]	[3]	[0]	[1]
1	[1]	[2]	[1, 6, 11]	[2, 5, 7]	[]	[]	[1]
2	[1, 2]	[2]	[0, 1, 6]	[2]	[]	[0]	[1]
3	[1, 2]	[2]	[0, 1, 2, 11]	[2]	[3]	[]	[1]
4	[]	[0, 2]	[1, 2, 6]	[2]	[3]	[]	[]
5	[1, 2]	[2]	[0, 1, 2, 6]	[2]	[3]	[]	[1]
6	[1, 2]	[2]	[1, 2]	[2, 5]	[]	[]	[1]
7	[1, 2]	[2]	[0, 1, 2, 6, 11]	[2, 5]	[]	[]	[1]
8	[1]	[2]	[0, 1, 2, 6]	[2]	[]	[0]	[1]
9	[1, 2]	[0, 2]	[1, 2, 11]	[2, 5, 7]	[]	[]	[1]

Figure 5: An example of reachability matrix R: in this data structure are visible 9 scenarios on the rows and 6 nodes on the column. For each cell a reachability list is contained.

the node number 4, its successors are nodes 3, 8, 6, 11, 10, 7. Considering, for example, the node 1, its predecessors is node 2.

3.3 Scenarios and reachability matrix R

The starting point of the influence study is the scenario, named ω that belongs to the ensemble Ω. It is defined as a possible realization of the first wave of individual's influence in the social network graph. They are generated randomly considering the same data structure but different arcs and weights, an example is visible in Figure 4.

Each scenario has the same possibility to be presented, so the scenario probability distribution is uniform. This is a strong assumption invoked to simplify the problem. From this consideration, the scenario probability is constant and equal to $\frac{1}{number\ of\ scenarios}$.
The reachability matrix R is a 3D data structure presenting on the rows the scenarios, on the columns the nodes composing the graph and in each cell the reachability list, that is composed by the influencing nodes incoming on the node in analysis, an example is shown in Figure 5. It is generated by iterating over all the graph nodes, checking their influencing predecessors and inserting them in a reachability list. An influencing hop is defined as if a randomly generated number between 0 and 1 is lower than the weight of the arc between the predecessors and considered node.

4 Mathematical model

The S-IMP is a stochastic problem formalized in Section 3 and with Gurobi, a tool that exploits the linear programming approach (LP), is computationally solved. It is a discrete problem, in which the feasibility region is a convex subset of solutions, determined by a group of constraints applied on the whole set of possible combinations of nodes that belong to the graph.
The variables used in the model are:

- $x_i^\omega \in \{0;1\}\,,\ i \in V,\ \omega \in \Omega;$
- $z_i^\omega \in \{0;1\}\,,\ i \in V,\ \omega \in \Omega.$

The constraints imposed on the feasibility region are:

	OF LP	Seed LP	Simu. time LP [s]	OF Heu	Seed Heu	Simu. time Heu [s]
Run 1	11.34	2;4	0.00898	11.96	2;4	0.000998
Run 2	11.94	2;4	0.0114	12.12	2;4	0.000996
Run 3	11.44	2;4	0.00998	12.90	2;4	0.000997
Run 4	11.88	2;4	0.00681	12.16	2;4	0.000996

Table 1: Results of the simulation with ad hoc graph model.

- The constraints on starting seed, it means that the number of nodes that compose the seed group must be lower or equal to the input size parameter K
$$\sum_{i \in V} z_i \leq K$$
- The reachability constraint for node i and scenario ω, instead, guarantees that node i can only be active in scenario w if at least one of the nodes in $R(\omega, i)$ is chosen as seed node
$$x_i^\omega \leq \sum_{j \in R(\omega;i)} z_j, \; i \in V, \; \omega \in \Omega$$

The objective function to be solved is

$$max \left\{ \sum_{\omega \in} p_\omega \cdot \sum_{i \in V} x_i^\omega \right\}$$

A parameter used in the model is the seed dimension K.

5 Design of the heuristic

A heuristic technique is any approach to problem-solving that uses a practical method or various shortcuts in order to produce solutions that may not be optimal but are sufficient given a limited timeframe or deadline. Heuristic methods are intended to be flexible and are used for quick decisions, especially when finding an optimal solution is either impossible or impractical and when working with complex data.
In this particular case, the heuristic method is based on the generation of the reachability matrix exposed in Subsection 3.3 and on the analysis of it in order to solve the objective function explained in Subsection 3.1. The goal of this approach is to obtain a comparable result with the exact LP method of Section 4, saving time and computational resources. After generating the social media graph, the first step is to extract the reachability matrix R of the scenarios , as described, after which for every scenario a frequency count of the nodes is performed. The basic idea is that a node present in many scenarios is a good candidate for the seed set. In order to respect the constraints in Section 4, only the first more popular K nodes are considered as the possible seed components z_l and they are next used for the objective function computation. The objective function is estimated as follows: for each scenario, the corrispective sub-seed is taken and the successors of those nodes are influenced. Then an average of those nodes of all the scenarios is computed. After the objective function computation, the final filtering is applied on all sub-seed sets, from which the most K popular nodes appearing over the different scenarios are chosen as the final seed set.

6 Performance

The heuristic and the LP approaches are compared both considering the objective function and the timing performances.

6.1 Timing performances

Initially, the ad hoc graph, exposed in Subsection 3.2, is tested. The parameter K for the seed size is set to 4, an influence seed composed by nodes $[2; 4]$ and eventually two other random nodes is expected. Doing different runs, both LP and heuristic always extracted only $[2; 4]$ as seed set. Other nodes are not taken into account since there are no significant profits in doing so, the whole results are reported in Table 1.

Once proven the S-IMP problem is working correctly, to further test and comparison between the LP and heuristic approach, a more complex graph is used, instantiating a scale-free network composed of 70 nodes, with K equal to 5 and 20 scenarios stochasticity. The problem is still a toy problem for a lower CPU time and a faster data collection. In order to have reliable results, 1000 runs are performed, obtaining Figure 6.

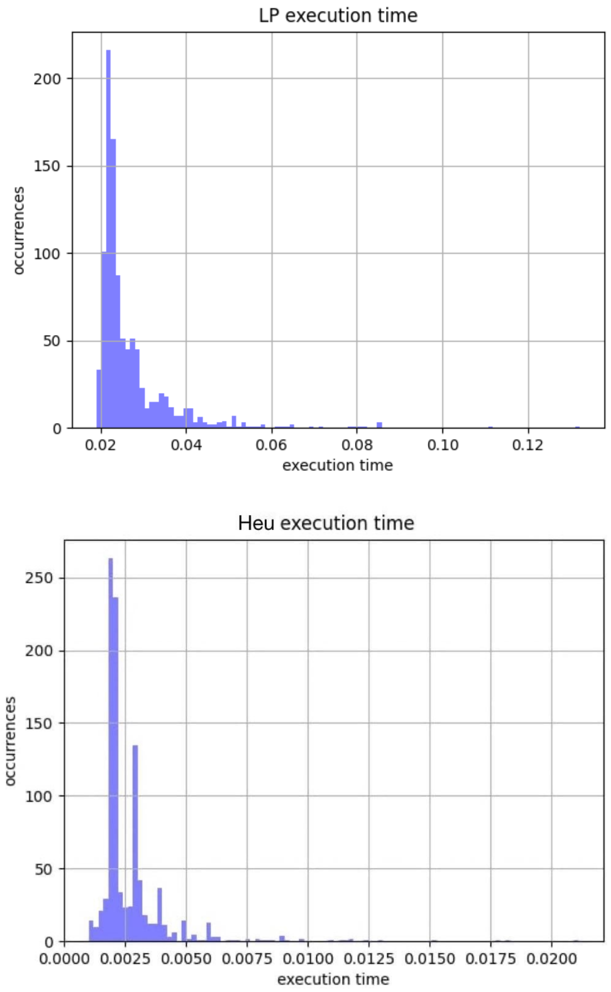


Figure 6: Above the LP execution time: $mean = 0.02697s, variance = 0.00976s$. Below instead the heuristic execution time: $mean = 0.00275s, variance = 0.00171s$.

	scale-free (500 n)	konect (800 n)
LP	0.04487	0.05286
Heurisitc	0.01696	0.01997

Table 2: Execution time for scale-free and konect graph. The LP timing increment is 37% greater with respect to heuristic variation.

From the plots it is visible that the slowest approach is the LP. The starting heuristic algorithm complexity is $O\left(N^3\right)$ and it is faster for small graphs while it becomes slower increasing the N value. To solve this issue, the Collections python library is exploited which uses sorting algorithms with a complexity $O\left(N \cdot logN\right)$. The execution time increases with the graph complexity. Considering the scale-free graph, with 500 nodes, or the konect graph, 800 nodes, the execution times are reported in Table 2.

6.2 Objective function analysis

In this subsection it is necessary to define the gap metric as the absolute value of the difference between the objective function results of the heuristic and of the linear programming approach:

$$gap = \frac{|OF_{LP} - OF_{heuristic}|}{OF_{LP}}$$

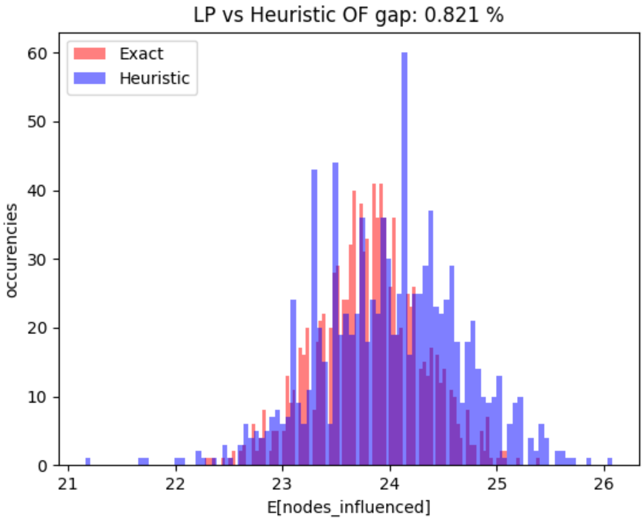


Figure 7: Occurrences of the average number of influenced nodes for each run. The two shapes are comparable, the gap value is 0.821%, so the heuristic results are very closer to the linear programming ones.

Using as reference a toy problem, 1000 runs of the whole analysis of a 70 nodes scale-free graph with 25 scenarios, the result distribution is obtained and shown in Figure 7.

6.2.1 In Sample Stability

After the preliminary data collection that confirmed the model correctness, a deeper analysis is performed. The first step is the In Sample Stability, a metric used to prove the model stability and the reliability of the solution. The aim of this analysis is to find a number of influenced individuals that is approximately constant, iterating on different realizations of the same graph, in which what is changed are the weights and the active nodes. In a more formal definition, the formula that has to be proved is

$$f(\hat{x_i};\tau_i) \approx f(\hat{x_j};\tau_j),\ \tau_m = \tau_0 + \tau_1 + ... + \tau_m$$

Each iteration is performed increasing the number of considered scenarios in order to track the model convergence and the variation around each objective function mean value. The analysis of the scale-free and of the konect graphs are reported in Figure 8 and 9. The higher the number of considered scenarios, the more stable is the model, with a converging mean value and a low uncertainty. The heuristic method has a little offset with respect to the exact mean value, but it is acceptable since it is about 3%. It is expected that with more runs it would converge to the exact value. For the konect data structure, the heuristic has a standard deviation twice the exact solver one, this result is probably due to the higher dimension of the graph. The important aspect is that the model presents a converging trend getting closer to the exact value of the objective function, that is about 60 influenced persons.

6.2.2 Out Sample Stability

The second step for the stability analysis is the Out of Sample Stability, again for both types of graphs. The test is divided into two stages:

1. Computation of the seed set on a reachability matrix every iteration bigger, with a larger number of scenarios.
2. For every seed set obtained, doing an influence expansion on 100 new scenarios that are not used in the previous stage. Working with scenarios that are never used before can lead to more fluctuating results, similarly when working with test and training dataset in Machine Learning, but it is important to fix the trend.

In Figure 10 and 11 the results are reported. In scale-free case the objective function results following a trend around 43 persons influenced. Those values are more fluctuating due to the usage, for the simulation, of a scenario set different from the one used for the creation of the seed.
Regarding the konect topology, increasing the number of scenarios, the solution tends to converge toward the mean value. The Heuristic again is more sparse due to the gap with respect to the exact seed values.

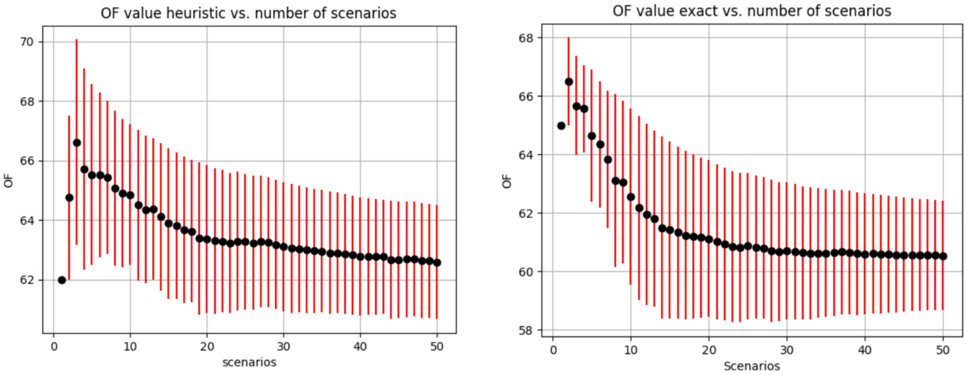


Figure 8: In Sample Stability scale-free graph: on the left the heuristic and on the right the LP.

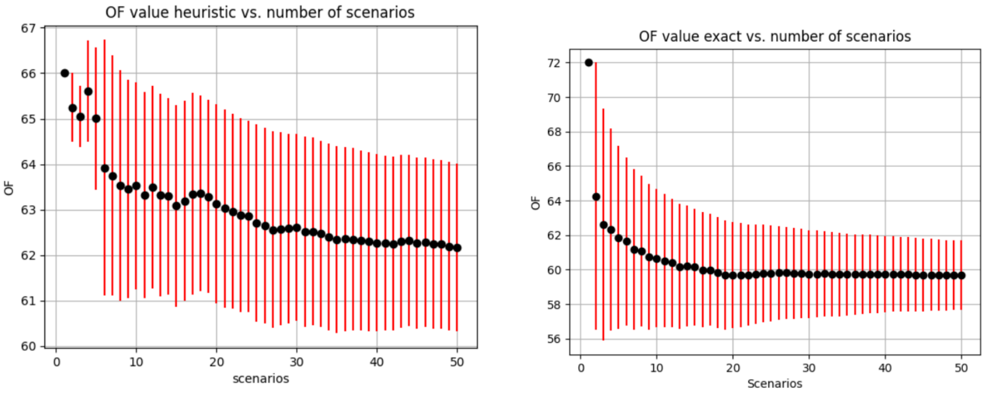


Figure 9: In Sample Stability konect graph: on the left the heuristic and on the right the LP.

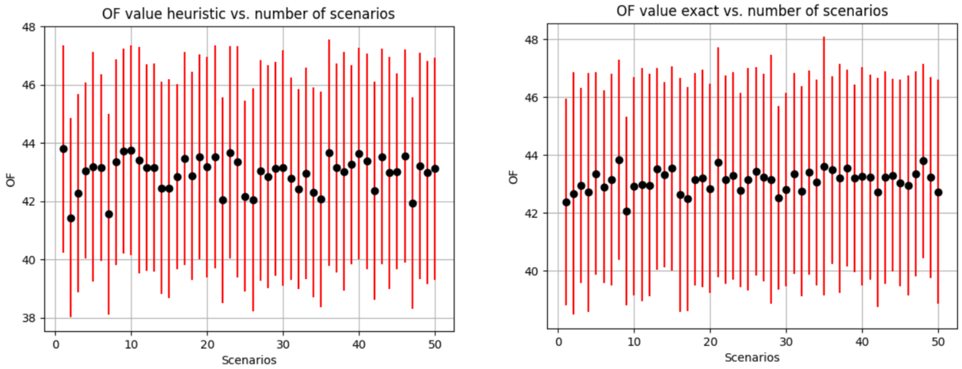


Figure 10: Out of Sample Stability scale-free graph: on the left the heuristic and on the right the LP.

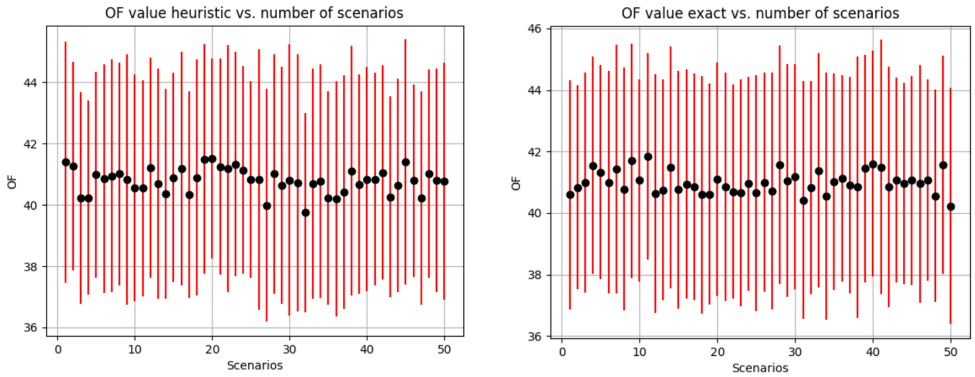


Figure 11: Out of Sample Stability konect graph: on the left the heuristic and on the right the LP.

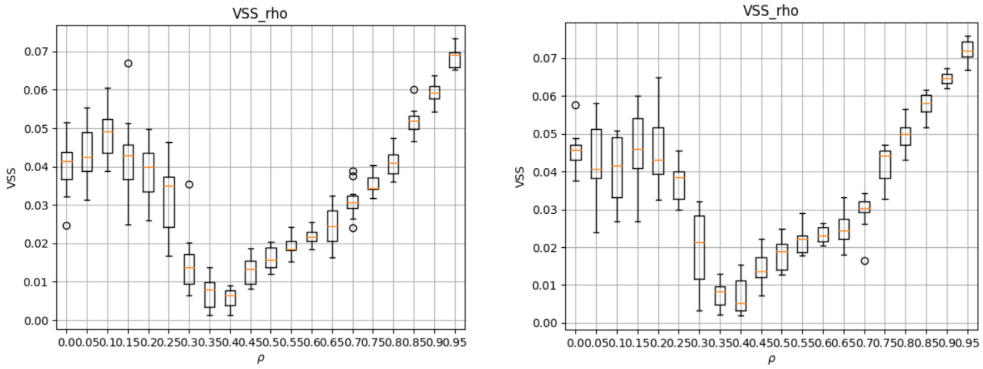


Figure 12: The VSS analysis: the results of the heuristic seed are compared with the one obtained from the S-IMP model by computing the difference. Each bar is a Gaussian shape seen by the top, smaller bar means lower variance.

The problem of the Out of Sample Stability is that it is not clear which is the correct mean to follow, it is necessary a correct reference value. For this purpose, a Vector Service Set (VSS) analysis is required. To obtain the VSS, a heuristic procedure is applied: different seed sets are obtained doing a frequency count and storing only the nodes with a frequency higher than a given threshold ρ. Then the mean set definition is used for an influence propagation simulation and the result is compared with the one obtained from the S-IMP model. The procedure is applied both to the scale free and konect graphs, and it is obtained that the best threshold value for the mean service set is 0.35 at which the difference between the means of the different trials is lower than 1%, confirming that the mean value fixed in the out of sample stability is correct. The resulting plots are shown in Figure 12.

7 Further improvement

To further improve the model and make it more complete two main aspects can be modified. One is the introduction of a cost for the seed selection: it is expected to spend more to hire a world famous influencer with respect to a local one. In this case, each node would have the cost attribute that can be proportional to the number of nodes it is connected to. The model would require an additional constraint

$$\sum_{j=0}^{K} c_j \cdot z_j \leq b, \; b = budget$$

The Second improvement that can be done is the scenario probability distribution. Instead of considering all the scenarios equally likely, it uses a metric to distinguish them, highlighting which one is more or less probable.

8 Conclusion

Finally the objective function results obtained by the linear programming approach and by the heuristic method are very close to each other. The timing is strongly in favor of the heuristic. In the comparison of the two algorithms an important role is taken by the data structure: the advantages of the heuristic or the better precision of the linear programming approach is clearly visible with complex graph with thousands of nodes.
A particular attention is paid on the concepts of reachability, predecessors and successors that can be generally defined as the relationship between nodes. In fact it is very important, to obtain a good results, a high interconnected network in which the Small World Effect can be exploited to achieve quickly the advertisement campaign purposes.
Nowadays social media marketing is a very good solution, in the future it is expected that it could be the mainly approach adopted by anyone who want to spread any kind of information through a specific or general set of population.

References

[1] Gruppi utenti e social media; https://business.trustedshops.it/blog/gruppi-utenti-social-media

[2] Esempi di social media marketing; https://www.nextredigital.it/esempi-social-media-marketing/

[3] Large-scale influence maximization via covering location, E. Güney, M. Leitner, M. Ruthmair, M. Sinnl, 17 May 2019

[4] A combined Deep-Learning and Transfer-Learning Approach for Supporting Social Influence Prediction, A. Cuzzocrea, C. K. Leung, D. Deng, J. J. Mai, F. Jiang, E. Fadda, November 2-5 2020

[5] Maximizing the Spread of Influence through a Social Network, D. Kempe, J. Kleinberg, E. Tardos

[6] Emergence of Scaling in Random Networks, Albert-László Barabási, Science 15 Oct 1999

[7] http://konect.cc/networks/librec-filmtrust-trust/