

Linear Programming for Influence Maximization Problems in Social Networks Based on Diffusion Models

Pengxuan Zhang*

Department of Mathematics, Wake Forest University, Winston-Salem, North Carolina, USA

*Corresponding author: zhanp21@wfu.edu

Abstract. Influence maximization is a challenging problem that combines optimization and the study of social sciences. The purpose is to find the most influential individual within a society with a limited number of initial nodes. With the increasing use of social networks these years, Understanding and using forceful diffusion will spread social events or innovations far more widely than the conventional perspective of communication. Many studies have focused on providing algorithmic methods, ignoring the study of finding the optimal point. Although it is still difficult for computer community to settle massive influence maximization issues computationally, it presents a significant opportunity to build operations research-based models, approaches, and analyses at this interface. In the current research, the influence maximization problem is investigated as a discrete optimization issue on social networks. It focuses on the creation of precise techniques for resolving massive cases near the optimal and will provide two classical diffusion models describing the propagation of influence in networks. Finally, given the additional constraints considered, an extension of the trigger model with effective analysis is proposed.

Keywords: Influence maximization; Social network; Diffusion models; Linear programming.

1. Introduction

As modern technology develops, online social networks have become an inevitable component of daily activities, including the frequent dissemination of information, influence, and ideas among members of society. Therefore, influential people are highly demanded by marketing companies and social organizations. There are many cases in which an idea or innovation begins to spread in the community. For instance, the use of a social mobile app among university students, breaking news on the Internet, or the idea of a political movement. Some will end fast while others last for an extremely long period. It is certain that for using the social spread wisely, we need to understand the underlying mechanism for such a dynamic model: To what extent people are affected by the key individuals in social relationships, such as the choice of their neighbors or friends? Those “word-to-mouth” referrals have fascinated many researchers, therefore have been a historical issue. A lot of related research has been done, such as the spread of innovation in different parts of the world or the study of viral marketing. They both take advantage of those social network effects. Therefore, a more general model, the *influence maximization problem* (IMP), has been proposed in the past few years and has been studied by Domingos and Richardson [1]. They built a fair descriptive model in a propagation setting to compute a maximizing set. Though it provides a beginning point for the study, the method is more heuristic rather than providing a specific solution to the theoretical optimal point. Though an optimal solution is accessible, the heuristic procedure does not prove the optimality.

Nowadays, modern linear programming solvers can find optimal solutions by computer those small-sized integer programs. In addition, solvers can be used for the heuristic solution by reporting the best feasible solution. Moreover, the duality of the linear programming (LP) problems enables the analysis of feasible solutions including analysis of deviations from the optimal solution, which is a huge process. This paper focuses more on operational models called *diffusion models* using LP.

2. Definitions

In this section, we will start by presenting a brief introduction to the LP and IMP problems. LP can be written in the form of $cx: Ax \leq b, x \geq 0$ and this can be solved through the simplex method and many others. To get integer values, the method of adding more constraints to ensure that the solution has only integer values is called Integer Linear Programming (ILP). These problems are typically regarded as NP-hard as well, meaning that heuristic methods must be used to solve moderate to large-sized problems. The branch and bound and cutting plane algorithms are two common methods for resolving ILP issues.

Branch-and-bound methods will use linear relaxation, which means ignoring the restriction for integer values. The steps will repeat in the restriction of certain bounds, and eliminate all the noninteger values. Therefore, when linear relaxation of the problem is solved, the integer value x_i^* will be chosen with two added child nodes to the tree, each representing additional constraints, say $x_i \leq \lfloor x_i^* \rfloor$ and $x_i \geq \lceil x_i^* \rceil$ influence maximization in the social network. There are different situations to do the pruning for solutions. The branch will be pruned for the following reasons: 1) there are no non-integer variables after the linear relaxation; 2) the node is infeasible; 3) the value of the node is worse than the value of the feasible solution of the best we already have; and 4) there is constraint. When all branches have been cut back, the procedure is complete, guaranteeing an ideal outcome. Another choice is to employ the concept of an approximative convex hull, which denotes the feasible region of the problem as well as the smallest convex set of the LP problem. Theoretically, further restrictions can be added to the existing function to force the corner points of the integer problem to only have these integer values, allowing the solution to also be applied to the integer issue.

Finding a good constraint to cut the linear relaxation program's optimal solution, then getting a new solution after adding the new cut, is another challenge if the solution contains non-integer values. Branch-and-cut algorithms, also known as branch-and-bound methods or branch-and-cut algorithms, are a recent development in exact solution procedures for integer systems.

We now turn to explain IMP. In IMP, a decision-maker seeks to strategically target seed persons over a social network at the lowest possible cost. As a result, each person has a unique influence on each of their neighbors. The issues have garnered a lot of interest because they have applications in a wide range of fields, including epidemiology, health care, marketing, and politics.

Three factors can be considered for IMP. The first is the number of selected seed nodes before the spread. Those nodes are the earliest to become active during the process, therefore they take advantage to influence others. The second is the final number of nodes in the spread. They are used to evaluate the effectiveness of the spread as well as the scope of influence. The last one is the total time of the process. The purpose is to enlarge the final node number while minimizing the nodes at the start for an unlimited period.

Given $G = (V, E, W)$ is a directed and weighted graph. In this network, $V = v_1, \dots, v_n$, which stands for the node set which represents the element of the social network. $E \subseteq V \times V$ is the number of edges belonging to the set, and W stands for weight. $w_{i,j}$ represents the weight from user i to user j in the network, and x_i^t of each user i ($i \in V$) will be observed when t is between 0 and t [2].

IMP is defined as by $\max \sigma(S): |S| = k; S \subseteq V$, a stochastic process of network, combined with optimization problem as above. The set of the node is represented as v that belongs to the seed set S . We note that it is a subset of them owning a fixed size k with $k \geq 1$, corresponding to the influential nodes. The overall aim is to recognize the seed nodes which are the most influential ones. Here the influenced and uninfluenced nodes are the same as active and inactive nodes.

3. Two Classic Diffusion Models

3.1. Basics of Diffusion Models

It is believed that the information spread will follow certain rules. Therefore, some well-known models are used and developed. In this part, the linear threshold (LT) model and the independent cascade (IC) model will be described and compared. The two components of IC time are rounds and node influences, which also serve to describe the likelihood of an impact. The activated nodes in the set will start to spread the messages to connected individuals with W as a weighted probability. However, in the LT model, each node is evaluated as an individual that has its threshold. Moreover, a node can only be affected when the outer influence is greater than its threshold. Though the two models are using different approaches, both are considered NP-hard problems.

To consider these two diffusion models in the social network, we will use *active* and *inactive* to differentiate the states of nodes. We assume nodes will become more active as well when connecting groups become active.

3.2. Linear Threshold Model

Sets, parameters, and decision variables used in models will be described in Table 1 for the basis of the mathematical formulation below. Models in Ref. [3] are used to illustrate the spread process in LT model.

Table 1. Sets, parameters, and variables used in the IT model

Sets	
I	Sets of nodes in the network
T	Periods
Parameters	
S_i	influence threshold of node i
f_{j_i}	influence of node j on node i
B	Constraint for initial seeds
b_i	Cost of selecting node i in the beginning
Decision variables	
x_{it}	Indicates whether node i is influenced at period t

The basic model is given in the following:

$$\max \sum_{i \in I} x_i |T|, \quad (1)$$

$$\sum_{i \in I} b_i |T| \leq B, \quad (2)$$

$$\sum_{j \in I_j} x_{j(t-1)} f_{j_i} \geq S_i x_{it} \quad i \in I, r \in T, \quad (3)$$

$$\sum_{j: i \in I_j} x_{j(t-1)} f_{j_i} \leq S_i + x_{it} \quad i \in I, r \in T, \quad (4)$$

$$x_{i(t-1)} \leq x_{it}, i \in I, r \in T, \quad (5)$$

$$x_{i(t-1)} \in \{0,1\}, i \in I, r \in T. \quad (6)$$

In the above function, the aim is for the number of influenced people to be maximized shown in Eq. (1). Eq. (2) restricts the magnitude of the seed set to k , such as the budget allocated to the initial seed set. The constraints of Eq. (3) make sure i fails to be affected at the time slit t if the complete

influence is below the threshold level of I . In contrast, Eq. (4) shows i will be influenced if the total influence exceeds the threshold. Eq. (5) ensures people will not change their status from active to the inactive one in the entire procedure. Eq. (6) stands for the usual binary restriction for the problem.

3.3. Independent Cascade Models

In this section, the IC model will be considered. When the node becomes active, it cannot be inactive anymore and will remain the status during the spread process like the LT model above. In the IC model, a node i is active in period t and can merely influence its neighbor once with P_{ij} as a probability, which can be reviewed as a random process like flipping a coin. If the attempt fails, node i cannot influence j anymore. When j is activated in period t , then the next step for i is to influence other neighbors in the $t + 1$ period. The process will continue until no more nodes can be activated. It is shown that for every pair (i, j) , a coin with P_{ij} is flipped at the beginning of the process, but the result will be stored and later be checked when i is activated and j is still inactive, as the time when the coin is flipped does not matter.

We used formulations and examples in Ref. [4] to present the IM problem in IC models.

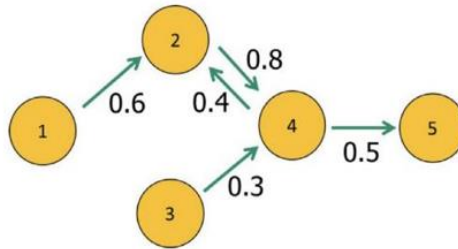


Fig. 1 A example social network in Figure 1 has five nodes and five arcs.

In Fig. 1, there is an example network with 5 nodes and 5 arcs. Here node 1 is selected as the beginning point. It has a probability of 0.6 to influence node 2, and if node 2 is successfully influenced, it will influence node 4 and so does node 5. Using the above notations and definitions, IC models can be constructed into integer linear programming models as shown in Table 2.

Table 2. Sets, parameters, decisions variables, and other notations

Sets	
S	the first batch of influencers
R	set of all potential outcomes
P_{ir}	group of close-by nodes that can affect a node
Parameters	
k	set of seeds
p_{ij}	The likelihood that node i will have an impact on node j
variables and the useful notations	
y_i	seed nodes
x_{ir}	Binary decision variable
z	objective function value for LP

The IC model is given in the following:

$$\max z = \sum_{r \in R} \sum_{i \in V} \mu_r x_{ir}, \quad (7)$$

$$s. t. \sum_{i \in V} y_i = k, \quad (8)$$

$$x_{ir} \leq \sum_{i \in P_{ir}} y_i, i \in v, r \in R, \quad (9)$$

$$0 \leq x_{ir} \leq 1, y_i \in \{0,1\}, i \in v, r \in R. \quad (10)$$

As suggested by Eq. (7), the IC model seeks to maximize the goal function like IT models for the determination of the anticipated amount of affected nodes. The size of the final seed set is constrained by Equation (8). The diffusion process is shown in Eq. (9). Node is only active when it is connected by active members or is a part of the seed set. Eq. (10) set the binary range for the variable, and forces y_i to be either 0 or 1. Then the binary requirements can be relaxed, and the optimal seed set y^* and corresponding maximal value z^* can be achieved. Many first active nodes may be present in a cyclic network solution without any of them being connected to the other first active nodes. This undesirable result may be avoided by identifying the set of predecessors as we have or by using a less efficient formula, but doing so requires the inclusion of a time index to the choice variables, which is not possible in the case of the formula we have used. In contrast to Sheldon's original formula [4], which exclusively tackles the acyclic issue, this approach is more practical because it takes into consideration both acyclic and cyclic circumstances.

3.4. Evaluation

It is well-known that the general LP problems belong to the NP-hard problem. Even if there is only one time period, which means the initial seed nodes will be the total affected nodes, the model is still known as NP-complete, so the simple version for the IM problem is still extremely hard to solve in polynomial time. Therefore, the heuristic approach should be applied to large-scale problems, such as the greedy approach and Period iteration approach.

According to research, IM has an ideal solution that is roughly within a factor of $(1 - 1/e - \varepsilon)$ in both two models [5]. Here ε is an infinitesimal increment and $e \approx 2.71828$, yielding a result that is marginally better than 63%.

4. Related Work

Many studies have recently created diffusion models based on the fundamental concepts outlined above in various disciplines according to needs [6]. For example, Chen and coworkers have extended the IC model to competitive social networks [7] where each node can either be in an active, inactive, or negative state. This is modeled using the Competitive Independent Cascade (CIC) framework. If a node is active in such a paradigm, it will either be in the positive or the negative state at the same moment. It will be in a negative condition if it is triggered by both positive and negative stimuli. When competing for the same target group, this paradigm is frequently utilized. Another study uses Benders decomposition algorithmic techniques to manage thousands of nodes and arcs in competitive IM based on the IC model in competition [8]. Similar to the original LT models, the competitive linear threshold model takes into account influences that can be either positive or negative and adds additional constraints [7]. The perfect target set selection (PTSS) problem is another issue with the LT model. The bare minimal initial set necessary to include every group member is used to study this topic. While using LTP to depict the PTSS, Ref. [9] suggests a combinatorial model for such a dynamic process. The combination of lower and upper bounds is frequently used to examine the minimum perfect target set. Based on ILP models, another challenge is also investigate, which is to identify the set of nodes that are impervious to influence. In their investigation of a robust optimization issue [10], Nannicini et al. assumed that the nodes' threshold and edge weights would fluctuate over a specified range and that time would not be a factor.

5. Conclusion

This paper reviews two basic diffusion models based on LP, and both of the models simulate the real-world scenarios of message spreading in social networks. They reflect the way the initial set of seed nodes influences connecting groups with change in time. The IT models require the total influence as least equal to the threshold of the current codes, while the IC model does not have such

restriction. Both models are considered NP-hard, therefore heuristic methods are required for providing solutions. In addition, further researches, including the extension of the basic models like competition models, is worth investigating as well. Those problems regarding LP could be in different forms, such as influence minimization, which contribute more to our understanding of the original IM problem.

6. References

- [1] Domingos P, Richardson M. Mining the network value of customers. Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. 2001, 57-66.
- [2] Wang R, Huang Y. Application of Linear Programming in Actual Production and Life. Science and Technology Innovation Herald, 2018, 9: 155-156.
- [3] Keskin M E, Güler M G. Influence maximization in social networks: an integer programming approach. Turkish Journal of Electrical Engineering and Computer Sciences, 2018, 26(6): 3383-3396.
- [4] Güney E. An efficient linear programming based method for the influence maximization problem in social networks. Information Sciences, 2019, 503: 589-605.
- [5] Sheldon D, Dilkina B, Elmachtoub A N, et al. Maximizing the spread of cascades using network design. arXiv:1203.3514.
- [6] Liu H. The More-for-less Paradox in the Linear Programming. Journal of China Three Gorges University, 1997, 15(1), 12-16.
- [7] Chen B L, Jiang W X, Chen Y X, et al. Influence blocking maximization on networks: Models, methods and applications. Physics Reports, 2022, 976: 1-54.
- [8] Kahr M, Leitner M, Ruthmair M, et al. Benders decomposition for competitive influence maximization in (social) networks. Omega, 2021, 100: 102264.
- [9] Ackerman E, Ben-Zwi O, Wolfowitz G. Combinatorial model and bounds for target set selection. Theoretical Computer Science, 2010, 411(44-46): 4017-4022.
- [10] Nannicini G, Sartor G, Traversi E, et al. An exact algorithm for robust influence maximization. Mathematical Programming, 2020, 183(1): 419-453.