






# Target Set Selection in Social Networks with Influence and Activation Thresholds

Zhecheng Qiang<sup>1</sup>(✉) , Eduardo L. Pasiliao<sup>2</sup> , and Qipeng P. Zheng<sup>1</sup> 

<sup>1</sup> Department of Industrial Engineering and Management Systems,  
University of Central Florida, Orlando, FL, USA

<sup>2</sup> Munitions Directorate, Air Force Research Laboratory, Shalimar, FL, USA

**Abstract.** Social media networks have gradually become the major venue for broadcasting and relaying information, thereafter making great influences in many aspects of our daily lives. With the mass adoption of the internet and mobile devices, social media users tend to follow and adopt their friends' or followers' thoughts and behaviors. Thus finding influential users in social media is crucial for many viral marketing, cybersecurity, politics, and safety-related applications. In this study, we address the problem through solving the influence and activation thresholds target set selection problem, which is to find the minimum number of seed nodes that influence all the users at time  $T$ . These time-indexed integer program models suffer from computational difficulties with binary variables at each time step. To this respect, this paper leverages computational algorithms, i.e., Graph Partition, Nodes Selection, and Greedy Algorithm to solve the models for large-scale networks. Computational results show that it is beneficial to apply the BFS Greedy algorithm for large scale networks. In addition, the results also indicate nodes selection methods perform better in the long-tailed networks.

**Keywords:** Networks · Integer programming · Target set selection · Greedy algorithm · Influence maximization

## 1 Introduction

Nowadays, the use of social media networks has become a necessary daily activity for people to interact with family and friends, access news, information and make decisions. Besides being a handy means for keeping in touch with friends and family, social media is more of a platform spreading the tremendous influence. Social media has great impact on businesses, politics, disease control and others. To this end, researchers have studied various practical problems in social media to better understand how the social media behaves and propagates the information. The problems include buzz prediction [7], volume prediction [24], infection prediction [3, 17], source prediction [21], link detection [19], target set selection [6, 8, 9, 14, 26] and firefighter problem [2].

## 1.1 Motivation and Problem Description

Here we focus on investigating the problem of target set selection. Formally, we define a social media network as a directed graph  $G = (V, E)$ , where users are defined as all nodes  $V$  and their friendships are defined as all edges  $E$ . Users are active when they repost the messages. Target set selection problem refers to find a minimum subset  $S \subset V$ , then all the nodes will be activated by  $S$ . The target set selection problem can be applied in the areas of viral marketing [11] and cyber security [4].

In the target set selection model, the seed nodes spread the influence until the diffusion process stops. The goal is to activate all the nodes. The influenced users refer to activated users who repost the messages. However, in reality, influenced users sometimes will repost the messages. But in most cases, even they are convinced or influenced by the message, users will not repost the message for certain reasons. In this case, influence not only refers to activation (repost) but also refers to the belief in the messages. Thus we build the influence and activation thresholds target set selection models to describe the situation. Here we introduce two thresholds, one is activation threshold ( $\phi$ ) and another one is influence threshold ( $\theta$ ). Users will be influenced first before be activated, thus we define  $\theta \leq \phi$ . The goal of the model is to influence all the nodes.

Our models are time-indexed integer program models, which can be divided into two parts, the first part is the information propagation. There are two widely used propagation models, namely Independent Cascade Model (IC) [14] and Linear Threshold Model (LT) [12]. Independent Cascade Model assumes every node has a single chance to activate its neighbors. In Linear Threshold Model, each node will be influenced by each neighbor according to a weight. When the total weights from its neighbors is larger than a threshold  $\phi$ , then the node will be activated. In this paper, we propose all the mathematical models based on the Linear Threshold Propagation Model. Here we set the weights as 1 for all the nodes. Thus an inactive node will become active if at least  $\phi$  of its neighbors are active in the previous step. The second part is the influence dominating part, which means the users should be either active or influenced through having at least  $\theta$  of activated neighbors at time  $T$ .

## 1.2 Literature Review

To investigate the influence and activation thresholds target set selection problem, the research of target set selection problem offers us some good insights. Kempe et al. [14] study the maximum active set (under the name target set selection) and show that it is NP-hard. They also provide a greedy algorithm within provable approximation guarantees based on the submodularity property of the objective function. Chen et al. [9] study the minimum target set selection problem and show the problem is hard to approximate within a poly-logarithmic factor. Besides, he comes up a polynomial-time algorithm to find an optimal solution when the underlying graph is a tree. Ackerman et al. [1] propose a combinatorial model for the minimum target set selection and prove

the combinatorial bounds for the perfect target set selection problem. Shakarian et al. [22] present a time-indexed formulation to find the minimum seeds for the target set selection and come up with a scalable heuristic based on the idea of shell decomposition. Spencer et al. [23] consider the problem of how to target individuals with subsidy in the network in order to promote pro-environmental behavior. It is also a target set selection problem and they use a time-indexed integer program formulation with as many time periods as the number of nodes in the network to tackle the problem. Günneç et al. [13] study the variation of the target set selection problem called least cost target set selection on social networks, and they propose greedy algorithm and dynamic programming algorithm to solve the problem for the tree structure network. Raghavan et al. [18] develop and implement a branch-and-cut approach to solve the weighted target set selection problem on arbitrary graphs.

### 1.3 Contribution and Organization

To our knowledge, the previous research involving target set selection focuses on the single threshold (activation threshold) target set selection. In this paper, we propose practical influence and activation thresholds target set selection mathematical model and its computational algorithms correspondingly.

The rest of the paper is organized as follows. Section 2 presents the novel Minimum Influence and Activation Thresholds Target Set Selection Model. In Sect. 3, we propose several computational algorithms for solving large scale networks. Section 4 shows the experimental results of the proposed model and its corresponding computational algorithms. Section 5 concludes the article.

## 2 Minimum Target Set Selection Model with Influence and Activation Thresholds

In this section, we introduce a time-indexed integer program to find the minimum number of influential seeds for the influence and activation thresholds target set selection problem. An artificial time index  $t$  taking values from 0 to  $T$  is introduced to model the order in which nodes become active. The messages could propagate at varying distances through different forms of social media. Cha et al. [5] observe that even for popular photos, only 19% of fans are more than 2 hops away from uploaders on Flickr.com. Ye et al. [25] find that, on Twitter, 37.1 percent message flows spread more than 3 hops away from the originators. Thus here we set  $T$  as 0,1,2,3, which means we only consider the cascades less than or equal to three time steps. The formulation uses a binary variable  $x_{i,t}$  to represent the status of node  $i$  at time  $t$ , which is 1 if node  $i$  is active at time  $t$  and 0 otherwise. Here  $\theta$  represents the influence threshold and  $\phi$  represents the activation threshold. Nodes should always be influenced before activated, so we set  $\theta \leq \phi$ .  $N(i)$  represents the neighborhood of node  $i$ . The Minimum Influential Seeds Model is as follows:

$$\min_{\mathbf{x}} \sum_{i \in V} x_{i,0} \tag{1}$$

$$s.t. \quad \theta x_{i,T} + \frac{1}{|N(i)|} \left\{ \sum_{j \in N(i)} x_{j,T} \right\} - \theta \geq 0 \quad \forall i \in V \tag{2}$$

$$\frac{1}{|N(i)|} \left\{ \sum_{j \in N(i)} x_{j,t} \right\} + \phi x_{i,t} \geq \phi x_{i,t+1} \quad \forall i \in V, t < T \tag{3}$$

$$\frac{1}{|N(i)|} \left\{ \sum_{j \in N(i)} x_{j,t} \right\} + \phi x_{i,t} \leq \phi - \epsilon + (1 + \epsilon)x_{i,t+1} \quad \forall i \in V, t < T \tag{4}$$

- $\theta \in (0, 1]$                       ← influence threshold
- $\phi \in (0, 1]$                       ← activation threshold
- $\theta \leq \phi$                               ← node is influenced before activated
- $N(i) = \{ j : (i, j) \in E \}$  ← Neighborhood, adjacent nodes

The objective function (1) aims to minimize the seed nodes activated at time 0. Constraints (2) are influential constraints, making sure that all the nodes should be either active or be influenced by at least  $\theta$  of active neighbors at time  $T$ . Constraints (3) refer that a node  $i$  will stay inactive at time  $t + 1$  when it is not activated at time  $t$ . Constraints (4) restrict that a node will stay active if it is originally active, which means when  $x_{i,t} = 1$ ,  $x_{i,t+1}$  should be 1 as well. In addition, the constraints make sure that a node will become active at time  $t + 1$  when it is activated at time  $t$ , which means when  $x_{i,t} = 0$ , the influence from its neighbors is larger or equal than  $\phi$ , then  $x_{i,t+1} = 1$ . Here we introduce two  $\epsilon$ , the first  $\epsilon$  restricts that node  $i$  should be active at time  $t + 1$  even if the influence from its' neighbors is  $\phi$ . The second  $\epsilon$  confirms that when  $x_{i,t} = 1$  and all the neighbors of  $i$  are active, the node  $i$  being active at time  $t + 1$  still holds.

### 3 Computational Algorithms

The time-indexed integer program model proposed in Sect. 2 is computationally intractable unless in very small instances because of the large number of binary variables. However, social media networks are usually in an extremely large scale. Thus we apply multiple computational algorithms to tackle the influence and activation thresholds target set selection model for larger scale networks in this manuscript. More details will be discussed in the rest of this section.

#### 3.1 Graph Partition

When the social media network is large-scale, solving the models exactly through Gurobi is very difficult. The most intuitive way is to solve multiple smaller sub-graphs instead of one large graph. Here we use techniques from Modularity and

Community Structure [10] in networks to divide the large graph into several smaller subgraphs. Then we solve the models exactly separately for each sub-graph.

### 3.2 Nodes Selection

When we're dealing with influence and activation thresholds target set selection problem for large-scale network, the large number of binary decision variables, which are  $|V||T|$  in total, makes the problem difficult to solve. In order to accelerate the computational speed, we could reduce the decision variables through adding some constraints to restrict that some of the nodes are not selected or some of the nodes should be selected. Here we come up with two methods, one is to delete the leaf nodes, and another is to choose the nodes with high degree.

**Leaf Nodes Deletion.** Leaf nodes in a connected graph may not be seeded because they'll influence or activate at most one neighbor directly. Thus we add the constraints (5) to remove the option of activating leaf nodes. In other words, all the leaf nodes will not be seeded using the method.

$$x_{i,0} + 1 \leq |N(i)| \tag{5}$$

**Degree Centrality Selection.** Nodes with high degree have more potential to influence and activate other nodes. Therefore, we assume the high degree nodes must be seeded. Here  $\frac{1}{|V|} \ll \rho < 1$  is defined as the criteria for choosing the seed nodes. When the total neighbors  $|N(i)|$  of node  $i$  is larger than  $\rho|V|$ , the node  $i$  will be seeded. Thus we add the constraints (6) to the original model in order to choose the nodes with more than  $\rho|V|$  neighbors as seed nodes.

$$x_{i,0} + \rho \geq \frac{|N(i)|}{|V|} \tag{6}$$

$$\begin{aligned} \rho = \epsilon &\text{ implies that } N(i) = 1 \text{ for a connected graph} \\ \rho = 1 - \epsilon &\text{ implies that } N(i) = V - 1 \end{aligned}$$

The larger the  $\rho$ , the nodes with higher degree will be selected as seed nodes. When  $\rho$  is  $\epsilon$ , the nodes having neighbors will all be selected. When  $\rho$  is  $1 - \epsilon$ , only the node connecting to all the other nodes will be selected.

### 3.3 Greedy Algorithm

We propose the greedy algorithm for the Minimum Influence and Activation Thresholds Target Set Selection problem. The greedy algorithm selects the seed node with the largest number of inactive neighbors in each iteration and adds it to the seed node set  $S$  until the stop conditions have been met. Then we update

the nodes threshold and active set in each iteration considering the propagation process. Here we update the threshold and active set based on the Breadth First Search (BFS).

For the BFS Search Greedy Algorithm shown in Algorithm 1, firstly we choose the seed node with the largest number of inactive neighbors and add it to the seed set  $S$ . Then we update the threshold and activation time step ( $T$ ) of an inactive neighbor node by adding the influence sent from the activated seed node. BFS search starts at the tree root and explores all of the neighbor nodes at the present depth prior to moving on to the nodes at the next depth level. Here we use a queue  $Q$  to store the parent nodes which will spread the influence within propagation step  $P$ .

---

**Algorithm 1.** BFS Search Greedy Algorithm
 

---

**Input:** Graph  $G = (V, E)$ , Propagation step  $P$

**Output:** Seed node set  $S$

```

1:  $A \leftarrow \emptyset, S \leftarrow \emptyset, Q$  be queue
2:  $threshold(i) = 0, T(i) = 0, ite(i) = 0$ 
3: while  $\exists v \in V \setminus A, n_A(v) < \theta deg(v)$  do
4:   Pick  $u \in V \setminus A$  with the most inactive neighbors
5:    $Q.enqueue(u)$ 
6:    $S = S \cup u$ 
7:    $A = A \cup u$ 
8:    $iteration = 0$ 
9:   while  $Q$  is not empty do
10:     $v = Q.dequeue()$ 
11:     $iteration = ite(v) + 1$ 
12:    for all  $\omega$  in  $N(v)$  do
13:      if  $\omega$  not in  $A$  then
14:         $threshold(\omega) = threshold(\omega) + \frac{1}{deg(\omega)}$ 
15:         $T(\omega) = max(T(\omega), iteration)$ 
16:        if  $threshold(\omega) \geq \phi$  then
17:           $A = A \cup \omega$ 
18:          if  $T(\omega) < P$  then
19:             $Q.enqueue(\omega)$ 
20:             $ite(\omega) = iteration$ 
21:          end if
22:        end if
23:      end if
24:    end for
25:  end while
26: end while
27: Return ( $S$ )

```

---

## 4 Computational Algorithms Comparison

In this subsection, we assess and draw comparisons between different computational algorithms introduced in the Sect. 3 for the minimum influence and activation thresholds target set selection model.

We consider a subset of real-life social networks as datasets in our experiment: Karate Club [27], Hamster Friendships Network [15], Facebook Network Dataset [16] and LastFM Social Network Dataset [20]. We set the time limit of 3600 s for bold methods in the following experiments. For the method of degree centrality, we set the  $\rho$  as 0.2.

The Karate network is a network of very small size and the result is shown in Table 1. For the Karate Club network, we could solve the model directly. However, the Leaf Node method could accelerate the computation slightly without sacrificing the performance. For larger size network of Hamster Dataset, the Graph Partition method couldn't generate a solution in one hour, so we don't include here in Table 2. For Leaf Node method, the model is not feasible which means we couldn't exclude all the leaf nodes as seed nodes for Minimum Influential Seeds model. We could see from the table the Original Model even performs better than the Degree Centrality method within the time limit of 3600 s. In addition, BFS Greedy method offers good solutions within much less time compared to the Original Model. The results of Facebook1 dataset are shown in Table 3. Facebook 1 is a dataset of low density. The Facebook 1 network has a large number of nodes with few friends. Thus it is easy for Gurobi to solve it directly. However, the Leaf node method has the shortest computation time for this dataset. The results of LastFM Asia Dataset are shown in Table 4, here Leaf Node performs the best compared with Original Model and Degree Centrality methods within the time limit of 3600 s.

**Table 1.** Karate network

Method	$\theta$	$\phi$	T	Seeded	Activated	Influenced	Obj	Time
Original Model	0.4	0.6	3	6	18	34	6	0.34
Graph Partition	0.4	0.6	3	7	32	34	7	0.09
Leaf Node	0.4	0.6	3	6	18	34	6	0.32
Degree Centrality	0.4	0.6	3	7	32	34	7	0.02
BFS Greedy	0.4	0.6	3	7	33	34	7	0.001

**Table 2.** Hamster dataset

Method	$\theta$	$\phi$	T	Seeded	Activated	Influenced	Obj	Time
Original Model	0.4	0.6	3	282	1543	1858	282	3600.47
Degree Centrality	0.4	0.6	3	291	1529	1858	291	3600.56
BFS Greedy	0.4	0.6	3	327	1766	1858	327	15.05

**Table 3.** Facebook 1 Dataset

Method	$\theta$	$\phi$	T	Seeded	Activated	Influenced	Obj	Time
Original Model	0.4	0.6	3	10	2888	2888	10	16.66
Graph Partition	0.4	0.6	3	10	2888	2888	10	20.17
Leaf Node	0.4	0.6	3	10	2888	2888	10	0.55
Degree Centrality	0.4	0.6	3	10	2888	2888	10	4.40
BFS Greedy	0.4	0.6	3	10	2888	2888	10	1.11

**Table 4.** LastFM Asia Dataset

Method	$\theta$	$\phi$	T	Seeded	Activated	Influenced	Obj	Time
<b>Original Model</b>	0.4	0.6	3	2850	6682	7624	2850	3601.25
<b>Leaf Node</b>	0.4	0.6	3	1498	6453	7624	1498	3601.52
<b>Degree Centrality</b>	0.4	0.6	3	3427	7624	7624	3427	3601.42
BFS Greedy	0.4	0.6	3	1675	7255	7624	1675	868.62

In summary, for the small size datasets, we could solve the problem directly using Gurobi. For the network of low density, especially when large portion of the nodes have few neighbors(long-tailed network), we could consider the Leaf Node method and Degree Centrality method. For the larger size datasets, normally the BFS Greedy will have better performance. The Graph Partition has poor performance and long computational time for the selected social media networks. For the Graph Partition method, it could result from the structure of network which is hard to divide into subgraphs. Furthermore, even it is divided properly, sometimes the size of the subgraph is still hard to solve directly.

## 5 Conclusion

The increasing popularity of social media networks has created the need for businesses, politicians and organizations to find influential users in social media to spread the influence. In this work, we have addressed the problem through developing the minimum influence and activation thresholds target set selection model. Our model allows us to find the minimum seed nodes that influence all the nodes at time  $T$ . In addition, we provide different computational algorithms to tackle the various datasets as well. They are Graph Partition, Leaf Node, Degree Centrality and BFS Greedy computational algorithms. Experiments in various datasets show that BFS Greedy is much more efficient than the other methods for large size datasets. Besides, leaf node deletion and degree centrality selection perform better in terms of long-tailed network.



## References

1. Ackerman, E., Ben-Zwi, O., Wolfowitz, G.: Combinatorial model and bounds for target set selection. *Theoret. Comput. Sci.* **411**(44–46), 4017–4022 (2010)
2. Anshelevich, E., Chakrabarty, D., Hate, A., Swamy, C.: Approximation algorithms for the firefighter problem: cuts over time and submodularity. In: Dong, Y., Du, D.-Z., Ibarra, O. (eds.) *ISAAC 2009*. LNCS, vol. 5878, pp. 974–983. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-10631-6\\_98](https://doi.org/10.1007/978-3-642-10631-6_98)
3. Bourigault, S., Lamprier, S., Gallinari, P.: Representation learning for information diffusion through social networks: an embedded cascade model. In: *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining. WSDM 2016*, pp. 573–582. ACM, New York (2016). <https://doi.org/10.1145/2835776.2835817>
4. Budak, C., Agrawal, D., El Abbadi, A.: Limiting the spread of misinformation in social networks. In: *Proceedings of the 20th International Conference on World Wide Web*, pp. 665–674. ACM (2011)
5. Cha, M., Mislove, A., Gummadi, K.P.: A measurement-driven analysis of information propagation in the Flickr social network. In: *Proceedings of the 18th International Conference on World Wide Web*, pp. 721–730 (2009)
6. Chen, C.-L., Pasiliao, E.L., Boginski, V.: A cutting plane method for least cost influence maximization. In: Chellappan, S., Choo, K.-K.R., Phan, N.H. (eds.) *CSoNet 2020*. LNCS, vol. 12575, pp. 499–511. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-66046-8\\_41](https://doi.org/10.1007/978-3-030-66046-8_41)
7. Chen, G.H., Nikolov, S., Shah, D.: A latent source model for nonparametric time series classification. In: *Advances in Neural Information Processing Systems*, pp. 1088–1096 (2013)
8. Chen, M., Zheng, Q.P., Boginski, V., Pasiliao, E.L.: Reinforcement learning in information cascades based on dynamic user behavior. In: Tagarelli, A., Tong, H. (eds.) *CSoNet 2019*. LNCS, vol. 11917, pp. 148–154. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-34980-6\\_17](https://doi.org/10.1007/978-3-030-34980-6_17)
9. Chen, N.: On the approximability of influence in social networks. *SIAM J. Discret. Math.* **23**(3), 1400–1415 (2009)
10. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. *Phys. Rev. E* **70**(6), 066111 (2004)
11. Domingos, P.: Mining social networks for viral marketing. *IEEE Intell. Syst.* **20**(1), 80–82 (2005)
12. Granovetter, M.: Threshold models of collective behavior. *Am. J. Sociol.* **83**(6), 1420–1443 (1978)
13. Günneç, D., Raghavan, S., Zhang, R.: Least-cost influence maximization on social networks. *INFORMS J. Comput.* **32**(2), 289–302 (2020)
14. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD 2003*, pp. 137–146. ACM, New York (2003). <https://doi.org/10.1145/956750.956769>
15. Kunegis, J.: KONECT - the Koblenz network collection. In: *Proceedings of the International Conference on World Wide Web Companion*, pp. 1343–1350 (2013). <http://dl.acm.org/citation.cfm?id=2488173>
16. Leskovec, J., Mcauley, J.J.: Learning to discover social circles in ego networks. In: *Advances in Neural Information Processing Systems*, pp. 539–547 (2012)

17. Qiang, Z., Pasiliao, E.L., Zheng, Q.P.: Model-based learning of information diffusion in social media networks. *Appl. Netw. Sci.* **4**(1), 111 (2019)
18. Raghavan, S., Zhang, R.: A branch-and-cut approach for the weighted target set selection problem on social networks. *INFORMS J. Optim.* **1**(4), 304–322 (2019)
19. Rodriguez, M.G., Balduzzi, D., Schölkopf, B.: Uncovering the temporal dynamics of diffusion networks. arXiv preprint [arXiv:1105.0697](https://arxiv.org/abs/1105.0697) (2011)
20. Rozemberczki, B., Sarkar, R.: Characteristic functions on graphs: birds of a feather, from statistical descriptors to parametric models (2020)
21. Shah, D., Zaman, T.: Detecting sources of computer viruses in networks: theory and experiment. *SIGMETRICS Perform. Eval. Rev.* **38**(1), 203–214 (2010). <https://doi.org/10.1145/1811099.1811063>
22. Shakarian, P., Eyre, S., Paulo, D.: A scalable heuristic for viral marketing under the tipping model. *Soc. Netw. Anal. Min.* **3**(4), 1225–1248 (2013). <https://doi.org/10.1007/s13278-013-0135-7>
23. Spencer, G., Howarth, R.: Maximizing the spread of stable influence: leveraging norm-driven moral-motivation for green behavior change in networks. arXiv preprint [arXiv:1309.6455](https://arxiv.org/abs/1309.6455) (2013)
24. Tsur, O., Rappoport, A.: What’s in a hashtag?: content based prediction of the spread of ideas in microblogging communities. In: *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, pp. 643–652. ACM (2012)
25. Ye, S., Wu, S.F.: Measuring message propagation and social influence on Twitter.com. In: Bolc, L., Makowski, M., Wierzbicki, A. (eds.) *SocInfo 2010*. LNCS, vol. 6430, pp. 216–231. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-16567-2\\_16](https://doi.org/10.1007/978-3-642-16567-2_16)
26. Yun, G., Zheng, Q.P., Boginski, V., Pasiliao, E.L.: Information network cascading and network re-construction with bounded rational user behaviors. In: Tagarelli, A., Tong, H. (eds.) *CSoNet 2019*. LNCS, vol. 11917, pp. 351–362. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-34980-6\\_37](https://doi.org/10.1007/978-3-030-34980-6_37)
27. Zachary, W.W.: An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**(4), 452–473 (1977)