15.093 Optimization Methods

Identifying Optimal Influence Strategy in Social Networks

Team Members: Arushi Jain (jarush08@mit.edu), Bibhabasu Das (bibha175@mit.edu)

INTRODUCTION

Social media is a powerful tool for influencing the masses and driving public opinion – be it political (Jan 6 US Capitol Riots), financial (GameStop short squeeze), medical (Covid-19), social (#MeToo, #BLM) or personal (targeted YT and IG ads). Consequently (and rightly so), many institutions aim to strategically target the right people to spread their messages through social networks either through personalized ads or through influencers.

PROBLEM STATEMENT

Given a network of people, how should one select optimal influencers from the network that can maximize the spread of our intended message/ campaign across the network?

Many studies[1] focus on developing and exploring the most influential initial nodes (or seeds) using greedy heuristics and numerical algorithmic methods. However, the solutions from an integer programming formulation can achieve a theoretical optimal point, which is often not possible with a greedy algorithm. Hence, in this project, we aim to develop an MIO approach to identify the optimal initial seeds in a social network for online campaigns.

The primary objective of the proposed work is to:

- Develop a mixed integer program to identify the optimal initial seeds in a directed weighted social network
- Develop a greedy heuristic baseline to understand the true impact of the MIO solution
- Perform comparative analysis with extensive numerical instances between the MIO and Greedy approaches
- Extend it to a use case which also considers negative opinion and harms the spread of the intended message

DATASET

We will use the Bitcoin OTC Stanford SNAP Dataset[2]. This is who-trusts-who network of people who trade using Bitcoin OTC Platform. The network has directed weighted edges where weight represents the amount of trust one node has on another. Table 1 shows the dataset statistics:

Dataset Statistics	
Nodes	6005
Edges	24,186
Range of edge weight	-10 to +10
Percentage of positive edges	93%

Table 1: Bitcoin OTC Dataset

This dataset is used to create a network where every node is considered to be a user and each edge highlights the effect it has on other users (both negative and positive) based on their directed edge weights obtained from the data.

EXPLORATORY DATA ANALYSIS

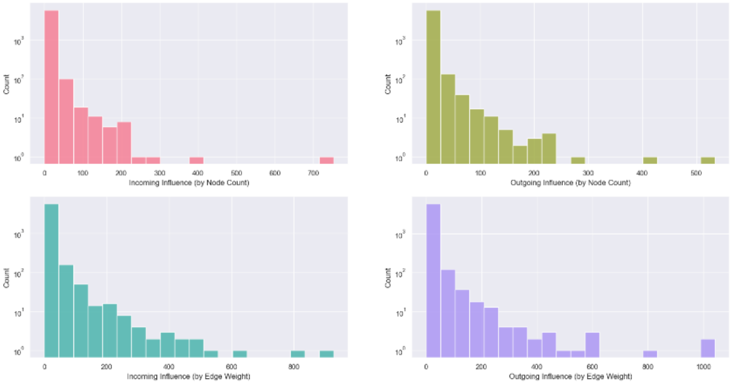

Figure 1: Distribution of Incoming and Outgoing Node Influence

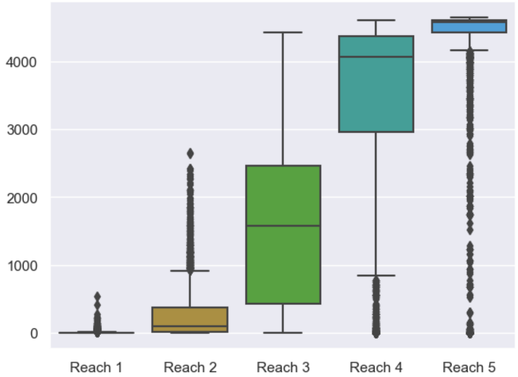

Figure 2: Distribution of Reach

To understand how the nodes are connected with each other in terms of count and edge weights, we analyze the following:

a. **Node Sway (Incoming Influence):** This represents how much influence other nodes have on a given node (incoming edges). We understand the distribution of the incoming influence in terms of the number of nodes directed to a given node as well as the incoming edge weights of the given node. The distributions are visualized on the left-half of Figure 1.

b. **Node Popularity (Outgoing Influence):** This represents how much influence a given node has on other nodes (outgoing edges). We understand the distribution of the outgoing influence in terms of the number of nodes a

[1] Güney, Evren. (2018). An Efficient Linear Programming Based Method for the Influence Maximization Problem in Social Networks.

[2] S. Kumar, F. Spezzano, V.S. Subrahmanian, C. Faloutsos. Edge Weight Prediction in Weighted Signed Networks. IEEE International Conference on Data Mining (ICDM), 2016

given node is directed to as well as the outgoing edge weights of the given node. The distributions are visualized on the right-half of Figure 1.

c. **Node Connectivity (Reach):** This represents the outgoing influence in terms of number of nodes at various reach levels. At reach level 1, a node can only influence its direct connections. At reach level 2, a node can influence its connections of connections and so on. This is a measure of "connectedness" of a given node and hence, its importance.

Observations:

- **Right-skewedness Popularity/ Sway graphs**: Most of the nodes are not very popular and not that easily swayed
- **Network Saturation**: After 5-levels, most of the nodes cover ~75% of the network

SYNTHETIC DATA

To select the target nodes for our campaigns, we need additional data in terms of influence threshold of each node (how easily a node will get influenced) and the cost of selecting that node as a seed. Since this data is not available for most of the publicly available datasets, we decided to synthesize this data. Intuitively, it should be designed such that the following conditions are satisfied:

- Influence Thresholds: Should be directly proportional to the number of "following" nodes
- Node Budgets: Should be directly proportional to the reach of the node in a network after certain levels

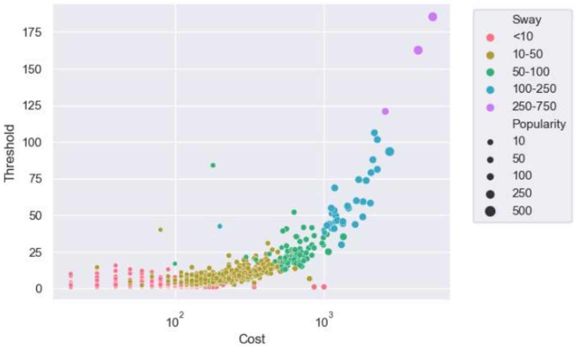

Figure 3: Comparison of Budgets, Threshold, Sway, and Popularity across Nodes

Thus, we decide influence thresholds for each node as 20% of the total incoming edge weight of that node. This means that for a given node, if the sum of the edge weights of its neighbour nodes that are influenced is more than 20% of the total incoming influence, we will consider the given node to be influenced.

We decide cost of each node as $10 per connected node at Reach level 2 with 20% noise. For example, if a given node is connected to 100 nodes at level 2, the cost of that node will lie between $800-$1200.

METHODOLOGY

INTUITION

Selecting the optimal set of influential nodes at the beginning of a campaign will yield a high expected number of activated nodes at the end. Each influenced node in the current time period will be likely to influence its follower nodes in the next time period. If the weighted sum of its influenced "following" in the previous time period exceeds the node influence threshold, the node will get influenced in the current time period. However, we will be constrained by our budget while selecting the initial nodes depending on their node costs. The number of time periods and the budget play a vital role in understanding and strategizing the impact we desire.

MIO FORMULATION

Through our MIO approach, we identified the individuals to be targeted in a network at the beginning of the campaign based on our budget in order to maximize influence after certain time periods.

The decision was whether a given node i will be influenced at time t. Since we were optimizing interventions for each node at each time period, we have exponential number of outcomes and hence, this problem was fit to be an MIO problem.

Sets	
I	Set of the nodes (people) of the network
T	Set of time periods (propagation duration)
I_i	Set of incoming neighbour nodes of node i
Parameters	
B	Budget allocated for selection of initial seeds
b_i	Cost of selecting node i in the initial seed
s_i	Influence threshold of node i
f_{ji}	Influence of node j on node i
Decision variables	
x_{it}	Indicates whether or not the node i is affected at period t

Table 2: MIO Formulation Details

The Primary Objective of the Influence Model:

$$\max \sum_{i \in I} x_{i|T|} \qquad (1)$$

The Associated Constraints of the Influence Model:

$$\sum_{i \in I} b_i x_{i0} \leq B \qquad i \in I \qquad (2)$$

$$\sum_{j: i \in I_j} x_{j(t-1)} f_{ji} \geq s_i x_{it} \qquad i \in I, t \in T/\{0\} \qquad (3)$$

$$\sum_{j: i \in I_j} x_{j(t-1)} f_{ji} \leq s_i + \sum_{j: i \in I_j} f_{ji} x_{it} \qquad i \in I, t \in T/\{0\} \qquad (4)$$

$$x_{i(t-1)} \leq x_{it} \qquad i \in I, t \in T/\{0\} \qquad (5)$$

$$x_{it} \in \{0,1\} \qquad i \in I, t \in T \qquad (6)$$

Here, our primary objective function in Eq. (1) aims to maximize the number of people influenced (active nodes) at the end of the planning horizon.

The budget constraint in Eq. (2) puts an upper limit on the amount that can be spent on the people that would be selected for the initial seed.

The flow constraint of Eq. (3) and Eq. (4) ensure that a node i is influenced if and only if the incoming edge weights of the influenced nodes exceeds its threshold. Particularly, from Eq. (3), if $x_{it} = 1$, then incoming flow to node i is greater than node threshold of node i at time t (else it is lower bounded by 0). From Eq. (4), if $x_{it} = 0$, then incoming flow to node i is less than node threshold of node i at time t (else it is upper bounded by total incoming flow to i)

The constraint of Eq. (5) ensures that once a person is influenced, they will always stay influenced as time progresses. Finally, the constraint of (6) ensures the usual binary restrictions.

GREEDY HEURISTIC

In order to compare our solution with some baseline, we also needed a heuristic-based solution which would likely be very fast but not that efficient. A naïve solution will select the seeds based on the maximum number of outgoing connections (popularity). However, such a node may have “weak” connections with its neighbours. Another solution can select the seeds based on the most “influential” nodes (nodes that have higher outgoing edge weights). We use a slight modification of this approach.

We used a metric of “Value”, which is basically the amount of influence a given node has per dollar cost of that node:

$$v_i = \frac{\sum_{j: i \in I_j} f_{ij}}{b_i} \qquad i \in I \qquad (7)$$

We then start selecting the nodes in increasing order of their “Value” until our budget is exceeded.

Algorithm 1: Greedy Heuristic to Select Seeds with budget B

```
compute $v_i$ for $i \in I$
sort nodes by decreasing $v_i$
selected_node_set = {}
while B is not zero and all nodes are not visited
        if $B \geq Cost(node_i)$
                add $node_i$ to selected_node_set
                $B = B - Cost(node_i)$
        visit next node
return selected_node_set
```

Based on the initial seed nodes, we compute the influenced nodes at each time period by checking if the total weight coming from the influenced nodes is greater than their respective thresholds.

Algorithm 2: Computing the Number of Influenced Nodes at the end of time period T

```
influenced_set = {}
influenced_matrix = n x n matrix of 0
for node in selected_node_set
        add outgoing influence of node to its followers in influenced_matrix
        add node to influenced_set
for time in [2, T]
        influenced_at_time_set = {}
        for all nodes
                if $sum(influenced\ matrix)\ at\ node\ column \geq Threshold(node)$
                        add node to influenced_set
                        add node to influenced_at_time_set
                for influenced_node in influenced_at_time_set
                        add outgoing influence of influenced_node to its followers in influenced_matrix
return  length(influenced_set)
```

RESULTS

The following graphs compare the spread of a message over 6 time periods with a budget of $100 for our MIO approach and the Greedy Heuristic (only "influenced" nodes are shown in the graphs)

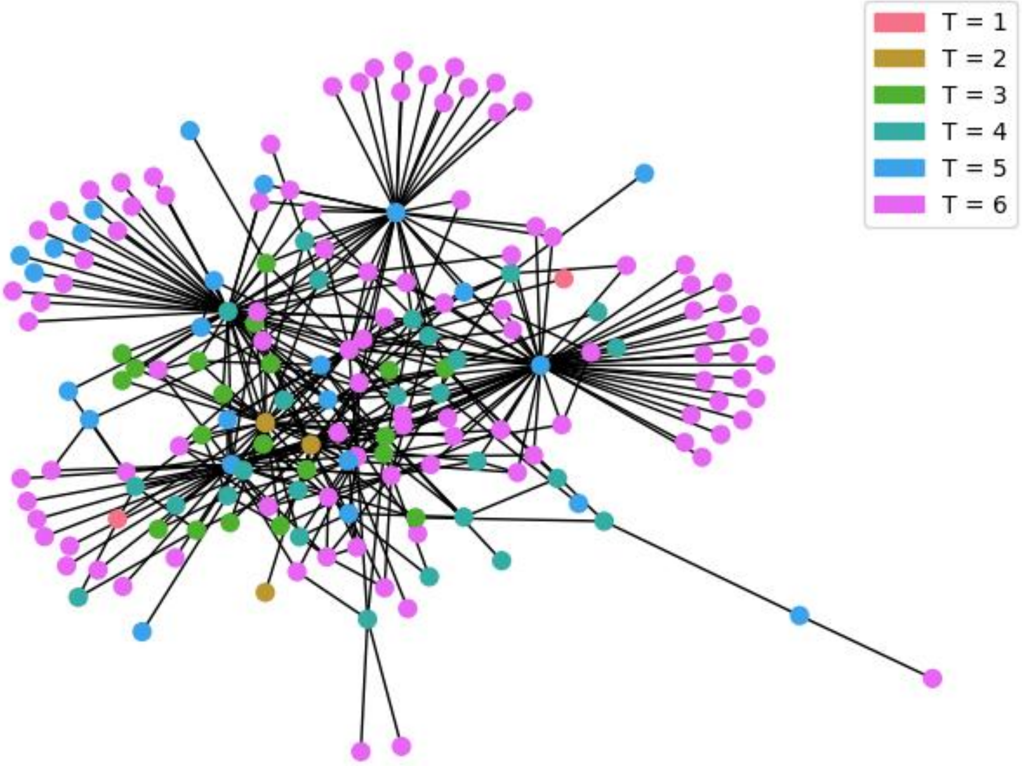

Figure 4: Spread of Influence in MIO at B = $100, T = 6 (178 Nodes)

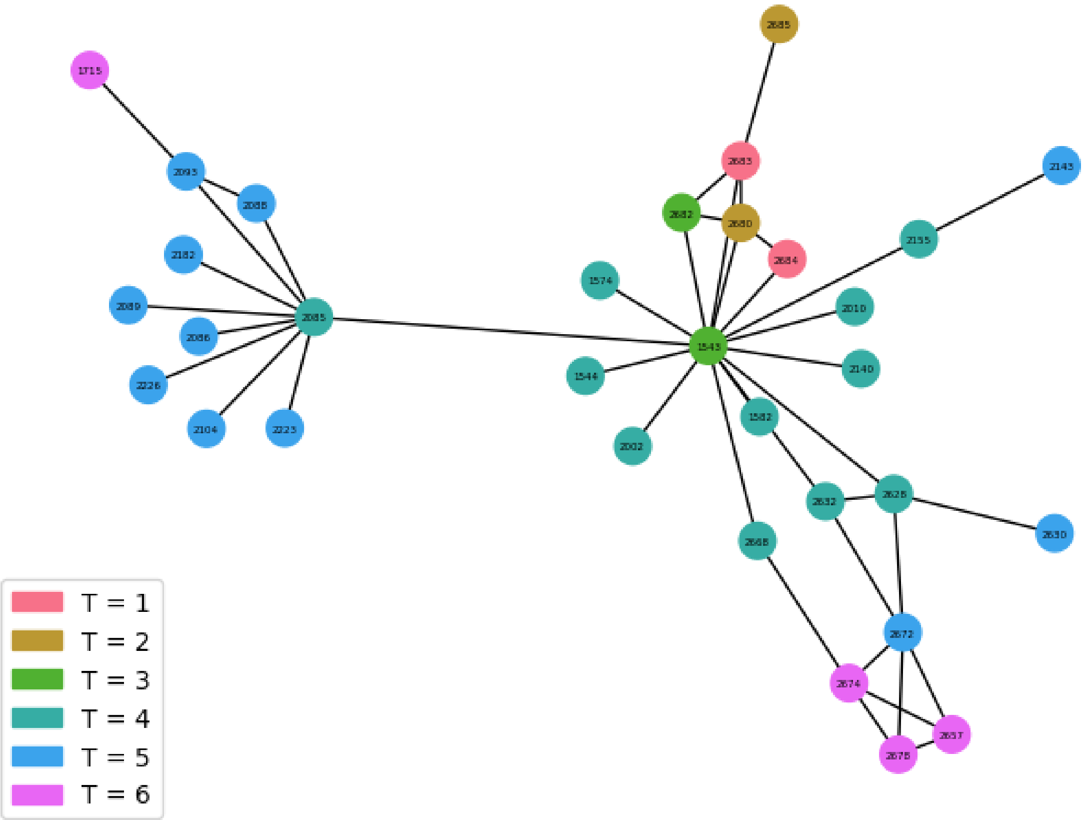

Figure 5: Spread of Influence in Greedy Heuristic at B = $100, T = 6 (32 Nodes)

We perform a more extensive comparative analysis of our MIO solution with the greedy solution with different seed budgets and influence propagation time periods.

Algorithm		B = $20	B = $50	B = $100	B = $500	B = $1000	B = $5000	B = $10000	B = $50000
T = 1	MIO	1	2	5	25	50	250	500	1836
	Greedy	1	1	2	7	16	54	84	318
T = 2	MIO	1	4	8	42	82	362	689	2667 (1%)
	Greedy	1	4	4	28	56	168	328	1658
T = 3	MIO	1	20	42	138	251	884 (30%)	1494 (43%)	4688 (7%)
	Greedy	1	9	6	30	93	274	584	2891
T = 4	MIO	1	32	77	245 (72%)	397(129%)	1001 (237%)	1808 (136%)	5710 (5%)
	Greedy	1	12	17	30	121	360	891	4453
T = 5	MIO	1	66	124	336 (400%)	496(521%)	1506 (294%)	2725 (82%)	6005
	Greedy	1	21	28	30	135	436	1545	4688
T = 6	MIO	1	72	178	428 (583%)	600(654%)	2116 (78%)	3383 (45%)	6005
	Greedy	1	25	32	30	136	522	2420	4712

Table 3: Number of Influenced Nodes (Values in Brackets show Optimality Gap)

Algorithm	B = $20	B = $50	B = $100	B = $500	B = $1000	B = $5000	B = $10000	B = $50000
MIO	20	50	100	500	1000	5000	10000	50000
Greedy	20	50	100	500	1000	4000	10000	49000

Table 4: Budget Used (MIO shows all constraints are always active)

Algorithm		B = $20	B = $50	B = $100	B = $500	B = $1000	B = $5000	B = $10000	B = $50000
T = 1	MIO	0.02	0.03	0.03	0.03	0.31	0.37	0.37	0.36
	Greedy	0.01	0.01	0.01	0.05	0.05	0.09	0.10	0.17
T = 2	MIO	12.61	16.26	21.93	19.20	24.51	26.02	29.17	1800
	Greedy	2.65	2.63	2.51	2.51	2.52	2.48	2.56	2.94
T = 3	MIO	26.23	31.72	36.68	77.53	202.65	1800	1800	1800
	Greedy	4.82	4.76	4.84	4.73	5.17	5.04	5.47	5.57
T = 4	MIO	60.25	65.45	67.32	1800	1800	1800	1800	1800
	Greedy	7.59	7.13	7.11	7.36	7.48	7.73	7.56	8.60
T = 5	MIO	88.44	100.26	1800	1800	1800	1800	1800	1648.39
	Greedy	10.94	10.74	10.51	10.26	10.48	10.81	10.73	12.02
T = 6	MIO	150.12	189.23	1800	1800	1800	1800	1800	764.34
	Greedy	12.32	12.41	13.27	13.37	12.71	13.03	14.09	14.94

Table 5: Computational Time in seconds (MIO had time limit of 1800 seconds)

INFERENCES:

1. **Greedy heuristic** is much faster but performs **extremely poorly compared to the MIO solution even at a time limit of 30 minutes and high optimality gaps.**
2. MIO solution saturates the network at B = $50,000 and T = 5 while the greedy solution saturates the network at B = $4,400,000 and T = 1 (not shown here). For reference, the sum of the costs for all nodes is also $4,400,000. The greedy solution is not able to saturate the network even at a budget of $4,000,000 (reaches only 87% of the network). **This means that the MIO solution requires 100x less budget than greedy heuristic to saturate the network.**
3. **At all budget and time combinations, the budget constraint in MIO is always active**. This ensures we are not wasting the allotted budget. On the other hand, we don't fully utilize the budget for the greedy approach at B = $5000 and B = $50000 (Note that the budget used in Greedy approach is independent of the time horizon)
4. Greedy heuristic, though intuitive, is sometimes unpredictable. For example, at T=3, greedy method influences a higher number of nodes at a budget of $50 compared to $100. This is explained via the table below:

	B = $50	B = $100
T = 1	2567	2682, 2683
T = 2	2565, 2566, 2567, 2568	2682, 2683, 2684, 2679
T = 3	2548, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572	1542, 2679, 2681, 2682, 2683, 2684

Table 6: Nodes Influenced for Greedy Heuristic

This happens because the greedy algorithm is able to select more "expensive" influencers at a higher budget which may have a better short-term reach. **Hence, it is possible to get a poorer solution at higher budgets with a greedy heuristic which is not the case with MIO.**

EXTENSION: WHAT ABOUT A NEGATIVE OPINION?

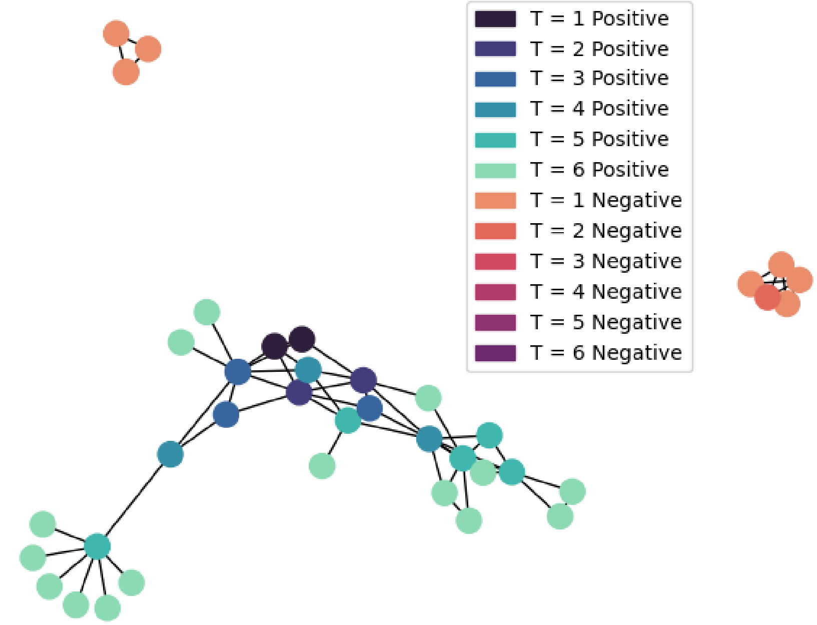

Figure 6: Spread of Influence in MIO at B = $100, T = 6
(30 Positive Nodes, 8 Negative Nodes)

Many a times, we have more than one opinion or products spreading or being marketed in social networks which can be quite polarizing and can harm the spread of our own intended message or product. We reformulate our model to account for an adversary whose marketing strategy harms our own.

We require the new parameters related to negative influence thresholds and the negative influence of one node on another. We synthesize the negative influence thresholds for each node as 5% of the total negative incoming edge weight of that node. To synthesize the negative influence of one node on other, we assume the same edge relationships but with different edge weights lying between 0 to 10.

MIO REFORMULATION

Additional Parameters	
r_i	Negative Influence threshold of node i
g_{ji}	Negative Influence of node j on node i
Additional Decision variables	
y_{it}	Indicates whether or not the node i is negatively affected at period t

Table 4: MIO Reformulation Details

The Primary Objective of the Negative Influence Model:

$$\max \sum_{i \in I} x_{i|T|} \qquad (8)$$

The Associated Constraints of the Negative Influence Model:

$$\begin{aligned}
&\sum_{i \in I} b_i x_{i0} \le B && && (9)\\
&\sum_{j: i \in I_j} x_{j(t-1)} f_{ji} \ge s_i x_{it} && i \in I, t \in T/\{0\} && (10)\\
&\sum_{j: i \in I_j} x_{j(t-1)} f_{ji} \le (s_i + (1-s_i) y_{it}) + \sum_{j: i \in I_j} f_{ji} x_{it} && i \in I, t \in T/\{0\} && (11)\\
&x_{i(t-1)} \le x_{it} && i \in I, t \in T/\{0\} && (12)\\
&\sum_{j: i \in I_j} y_{j(t-1)} g_{ji} \ge r_i y_{it} && i \in I, t \in T/\{0\} && (13)\\
&\sum_{j: i \in I_j} y_{j(t-1)} g_{ji} \le (r_i + (1-r_i) x_{it}) + \sum_{j: i \in I_j} g_{ji} y_{it} && i \in I, t \in T/\{0\} && (14)\\
&y_{i(t-1)} \le y_{it} && i \in I, t \in T/\{0\} && (15)\\
&x_{it} + y_{it} \le 1 && i \in I, t \in T/\{0\} && (16)\\
&x_{it}, y_{it} \in \{0,1\} && i \in I, t \in T && (17)
\end{aligned}$$

We now have a new decision variable y_{it} which represents if node i is *negatively* influenced at time t. The new constraints are accented in Blue. The constraint in Eq. (11) is similar to the constraint in Eq. (4). If both x_{it} and y_{it} are 0, then the incoming flow to node i is less than node threshold of node i at time t. If $x_{it} = 1$ and $y_{it} = 0$, then it is upper bounded by the sum of its threshold and the total incoming flow to i (implies free) . If $x_{it} = 0$ and $y_{it} = 1$, then it is upper bounded by 1 since we do not have any positive influence. The scenario $x_{it} = 1$ and $y_{it} = 1$ is not possible due to Eq. (16) which ensures that a node cannot be both positively or negatively influenced. Constraints in Eq. (13)-(15) are analogous to constraints in Eq. (10)-(12) for the negatively influenced nodes.

RESULTS

Figure 6 shows the result for the negative opinion model for B = $100, T = 6. Compared to a basic model where we could influence 178 nodes in our favour, we are able to influence only 30 nodes in the presence of a competitor.

CONCLUSION AND FUTURE WORK

We implemented an MIO solution to maximize influence in online social networks and compared it against a greedy heuristic over extensive numerical instances. We observed that MIO significantly improves the reach over the network over a greedy solution at the cost of higher computational times. We concluded that the MIO technique never wastes the budget and requires 10x less budget than the greedy solution to reach the entire network. We also implemented a variant of this model with a competitive / negative opinion and found that it greatly affects the reach with the same time horizon and budget constraints. This solution can be extremely helpful for organizations that are looking to promote their products, services, and ideas through online influencer marketing by targeting the right influencers given by our model.

As a next step, we could analyze the negative opinion variant over different time horizons and budget constraints and see how our reach over the network scales. We can also add further complexities like influence dynamic thresholds for nodes with respect to time and budget. Additionally, considering that the social network connections are ever evolving, robust optimization based on the edge connections and weights can also be an interesting trajectory to explore.