GUILHERME CARBONARI BONETI

MAXIMIZING INFLUENCE BLOCKING WITH COMPETING CASCADES USING INTEGER LINEAR PROGRAMMING

(*pre-defense version, compiled at June 27, 2024*)

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da Computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Ciência da Computação*.

Orientador: André Luís Vignatti.

CURITIBA PR

2024

RESUMO

No problema de *maximização de bloqueio de influência* nosso objetivo é minimizar a propagação de desinformação em uma rede. Mais especificamente, dado um conjunto de nós que iniciam a propagação de desinformação e um inteiro k, devemos encontrar k nós na rede para servir como ponto de partida para informações concorrentes (digamos, corretas) de forma a minimizar a propagação de desinformação. Neste trabalho, propomos uma versão do problema sob um modelo de disseminação derivado do modelo competitivo de limiar linear. Apresentamos uma formulação de programação linear inteira para o problema que, até onde sabemos, inaugura o uso de programação matemática em problemas de disseminação com duas cascatas concorrentes. Realizamos experimentos para verificar a qualidade de nossa formulação, avaliando o gap de integralidade e a escalabilidade. Tais resultados podem servir como referência para futuras soluções usando programação linear inteira. Os resultados desse estudo foram publicados nos Anais do Simpósio Brasileiro de Pesquisa Operacional (SBPO) de 2023.

Palavras-chave: Maximização de Bloqueio de Influência. Desinformação. Programação Linear Inteira.

ABSTRACT

In the *influence blocking maximization* problem, we aim to minimize the spread of disinformation in a network. More specifically, given a set of nodes that initiate the spread of disinformation and an integer k, we must find k nodes in the network to serve as the starting point for competing (say, correct) information so that the misinformation spread is minimized. In this work, we propose a version of the problem under a dissemination model derived from the Competitive Linear Threshold model. We present an integer linear programming formulation for the problem that, as far as we know, inaugurates the use of mathematical programming in dissemination problems with two competing cascades. We performed experiments to verify the quality of our formulation, evaluating the integrality gap and the scalability. Such results can serve as a baseline for future solutions using integer linear programming. The results of this study were published in the Proceedings of the Brazilian Symposium on Operations Research (SBPO) of 2023.

Keywords: Influence Blocking Maximization. Misinformation. Integer Linear Programming.

CONTENTS

1 INTRODUCTION

In recent years, there has been a growing interest in the propagation of influence through social networks, owing to the rapid growth of social media and their increasing importance in our society. In these environments, any piece of information can be disseminated rapidly and achieve significant scale. Unfortunately, this has also led to the proliferation of misinformation on social media platforms.

Misinformation has always been a problem, but it seems to have gained momentum with the rise of social media (Lazer et al., 2018). People tend to believe information that aligns with their social narratives and discredit information that challenges their beliefs (Lewandowsky et al., 2012). Social media's structure and methods of spreading information can amplify the spread of misinformation. (Vosoughi et al., 2018) found that on Twitter, fake news is 70% more likely to be shared than real news. Several recent studies have demonstrated the potential for misinformation to influence societal behavior. For example, Allcott and Gentzkow (Allcott and Gentzkow, 2017) analyzed the potential impact of misinformation on the outcome of the 2016 United States presidential election. Another instance is the proliferation of questionable sources on major social media platforms during the COVID-19 outbreak, as documented by Cinelli et al (Cinelli et al., 2020).

1.1 CONTRIBUTIONS

We formalize and investigate a problem motivated by the issue of misinformation, which we call *Influence Blocking Maximization in the SCLT model*, drawing on previous research on this subject (He et al., 2012). This problem arises when two information flows compete for dominance in a network, and our approach can help identify optimal strategies for blocking or mitigating the impact of unwanted information flows.

Our contribution comprises the formal definition of a version of the problem based on the SCLT dissemination model, the proof that this problem is NP-Hard, an integer linear programming formulation for the problem, and experiments evaluating the *integrality gap* and running time to verify the quality of the proposed formulation. As far as we know, this is the first ILP formulation for the influence blocking class of problems. Previous ILP formulations address problems with a single information flow (Fischetti et al., 2018; Ghayour et al., 2019), but do not consider the two opposite flows that occurs when modeling misinformation situations. Our results demonstrate that our formulation performs closely to its relaxed version, indicating its efficiency and effectiveness. This research was published in the Proceedings of the Brazilian Operational Research Symposium (SBPO) 2023 (Boneti et al., 2023).

1.1.1 Organization

The rest of this work is organized as follows. Section 2 contains the related work. Section 3 presents the diffusion models and the model we call the Simplified Competitive Linear Threshold Model, which we use in this work. Section 4 presents the problem definition we deal with, detailing the process of deriving it from the problem definition originally presented by (He et al., 2012) and the proof of its NP-hardness. Section 5 presents the formulation of the integer linear program that models the problem, and the proof of the correctness of such formulation. Section 6 describes the computational experiments performed, and discusses the results. We conclude the work in Section 7.

2 RELATED WORKS

Various models have been developed to capture the characteristics of real-world information propagation. Among them, the independent cascade and linear threshold models, originally described by (Kempe et al., 2003), are well-known. To address influence propagation problems under these models, several techniques have been proposed. Approximation algorithms, such as CELF (Leskovec et al., 2007) and CELF++ (Goyal et al., 2011), exploit submodularity to find near-optimal node selections for the Influence Maximization problem. To significantly reduce CELF's runtime, (de Melo, 2021) develops a preprocessing heuristic algorithm. Mathematical programming methods were used in (Fischetti et al., 2018) and (Ghayour et al., 2019). The former introduces the Generalized Least Cost Influence Propagation (GLCIP) and an integer linear programming (ILP) for this problem, while the latter proposes ILP formulations to handle the Influence Maximization (IM) and Target Set Selection (TSS) problems. (He et al., 2012) introduced a new problem called Influence Blocking Maximization (IBM), which was studied under the Competitive Linear Threshold (CLT) model, considering two competing information flows within the network. The authors developed a heuristic algorithm to address the IBM problem.

While there are other studies that tackle the IBM problem, none of them present an ILP approach to address it with competing cascades. Therefore, in this study, we introduce what we believe to be the first ILP formulation for this problem. Given the different variations of the IBM problem that have been studied, we believe that our work can serve as a valuable foundation for developing ILP models to tackle these variations.

3 DIFFUSION MODELS

In real world situations, information is transmitted through a variety of channels, including social media, television, and face-to-face conversations. The precise way by which a particular idea spreads and the factors that drive individuals to accept or reject it can be difficult to discern. Nevertheless, diffusion models aim to capture the dynamics of information propagation in social networks, providing valuable mathematical frameworks for studying influence dissemination, despite limitations in capturing real-world nuances. Next, we present two important diffusion models and a simplified version we use for our model.

3.1 LINEAR THRESHOLD MODEL

The Linear Threshold (LT) Model, as defined by (Kempe et al., 2003), models a social network as a directed graph $G = (V, E)$, where V represents the set of vertices (individuals) and E represents the set of edges (connections) between them. Each edge (u, v) in the graph connecting vertices u and v is assigned a weight $w_{u,v}$. Additionally, each vertex v is assigned a threshold value, t_v, that is chosen uniformly at random from the interval $[0, 1]$. In this model, a vertex, or node, can either be active or inactive, depending on whether it has been influenced by information or not, respectively. In this model, a node v is considered influenced if the sum of the weights of its incoming neighbors surpasses its threshold, i.e.,

$$\sum_{\forall (u,v) \in E} w_{u,v} > t_v.$$

The diffusion process starts with an initial set of nodes carrying the information. At each discrete time step, an inactive node becomes active if the sum of the weights of its active in-neighbors exceeds its threshold. The diffusion process continues until the entire network is influenced.

3.2 COMPETITIVE LINEAR THRESHOLD MODEL

The Competitive Linear Threshold (CLT) Model is an extension of the LT model, which considers two competing ideas for influence in the network, represented by positive and negative diffusion, as proposed in (He et al., 2012). Once a node is influenced by either diffusion, it cannot change the value of its activation (positive or negative) and be activated by the opposite behavior. Similar to the LT model, in CLT we have negative and positive seed sets, denoted by S^- and S^+, respectively. Additionally, each node in the network has positive and negative thresholds, denoted by t_v^- and t_v^+, and each edge has positive and negative weights, denoted by w_v^- and w_v^+.

At each time step, negative and positive influences propagate independently using negative thresholds and weights, and positive thresholds and weights, respectively. Since a node can only be activated by one diffusion, if both negative and positive thresholds are exceeded, the negative diffusion is assumed to win, and the node is negatively activated.

3.3 SIMPLIFIED COMPETITIVE LINEAR THRESHOLD MODEL

In this work, we use a simplified version of the Competitive Linear Threshold (CLT) model, which we refer to as the Simplified Competitive Linear Threshold (SCLT) model. Let $\delta_{\text{in}}(v)$ be the number of in-neighbors of a node v. In the SCLT model, we set both the positive and negative thresholds to half of the number of incoming neighbors for each node, i.e., $t_v^- = t_v^+ = \lfloor \delta_{\text{in}}(v)/2 \rfloor$. Such value is often referred to as the majority threshold (Chen, 2009). Moreover, all edge weights are assigned a fixed value of 1. This simplification allows us to focus on the fundamental aspects of the diffusion model.

4 PROBLEM DEFINITION

In this section, we define the problem we deal with in this work, called the *Influence Blocking Maximization in the SCLT model*. However, before we define our problem, we present the definition of the original problem under the CLT model, as done by (He et al., 2012). The change is only in the diffusion model, from CLT to SCLT, but this changes the definition of the problem, so that the simplest way to understand how this change occurs is by presenting the original problem first, and only then pointing out the modifications.

Let S^-_{end} be the set of negative nodes at the end of the diffusion process, and $s^-_{\text{end}} = |S^-_{\text{end}}|$. In the CLT model, as presented by (He et al., 2012), the thresholds values are set according to a probability distribution. Thus, the value s^-_{end} is a random variable. Clearly, s^-_{end} depends on the fixed input data and also the chosen solution S^+, which is the only variable. Thus, our notation only makes explicit the dependency on S^+, and we write $\Pr\left(s^-_{\text{end}} = \ell \,\middle|\, S^+\right)$ to denote the conditional probability that the set S^-_{end} has size ℓ given that the set S^+ was chosen as the solution. Given a solution S^+, the expected size of the negative nodes s^-_{end} is,

$$\mathbb{E}\left[s^-_{\text{end}} \mid S^+\right] = \sum_{\ell=0}^{|V|} \ell \cdot \Pr\left(s^-_{\text{end}} = \ell \,\middle|\, S^+\right).$$

To measure the impact of a solution S^+, (He et al., 2012) consider the difference between two scenarios, when the solution is indeed S^+, and when the solution set is empty (i.e., letting the negative spread without blocking it). This is called the *expected blocked negative influence* of S^+, and is formally defined as

$$\sigma(S^+) = \mathbb{E}\left[s^-_{\text{end}} \mid \{\emptyset\}\right] - \mathbb{E}\left[s^-_{\text{end}} \mid S^+\right],$$

and we want to maximize this quantity. We can now define the problem, as presented by (He et al., 2012).

> **Problem** (Influence Blocking Maximization in the CLT model (He et al., 2012))**.**
> **Input:** graph $G = (V, E)$ with thresholds t_v^+ and t_v^- for each $v \in V$, weights w^+ and w^-, a negative seed set S^-, and a positive integer k.
> **Output:** a positive set of nodes S^+ of size k such that maximizes $\sigma(S^+)$.

It has been proven by (He et al., 2012) that IBM is NP-hard under the CLT model.

In this work we consider the same problem, but in the SCLT model. This causes some modifications to the original definition. First, the positive and negative thresholds are the same, so we can use a unified t_v notation. Also, the thresholds are fixed at $\lfloor \delta_{\text{in}}(v)/2 \rfloor$. This means that the optimization function become deterministic rather than probabilistic. Thus, we replace the mathematical expectation with deterministic functions $s^-_{\text{end}}(\{\emptyset\})$ and $s^-_{\text{end}}(S^+)$. Note that $s^-_{\text{end}}(\{\emptyset\})$ becomes a fixed deterministic value, which can be computed in polynomial time (just simulate the diffusion process without selecting positive nodes for the solution). So, if before the objective was to maximize the difference $s^-_{\text{end}}(\{\emptyset\}) - s^-_{\text{end}}(S^+)$ but the first term is fixed, now it is the same as maximizing $-s^-_{\text{end}}(S^+)$. Another equivalent way is to say that we want to minimize $s^-_{\text{end}}(S^+)$, i.e., the objective is to choose S^+ in order to minimize the negative spread. Now we are ready to define the problem we deal with in this work.

> **Problem** (Influence Blocking Maximization in the SCLT model)**.**
> **Input:** graph $G = (V, E)$ with thresholds $t_v = \lfloor \delta_{\text{in}}(v)/2 \rfloor$ for each $v \in V$, a negative seed set S^-, and a positive integer k.
> **Output:** a positive set of nodes S^+ of size k that minimizes the negative spread.

In the following, we show that although the IBM in the SCLT model is simpler than its version in the CLT model, it remains NP-Hard. To do this, we present a reduction based on the well-known NP-Complete Dominating Set problem (Garey and Johnson, 1979). In the Dominating Set problem (DS), we are given an undirected graph $G = (V, E)$ and a positive integer $k \leq |V|$. The problem asks whether there exists a set S of size at most k in G such that every vertex in the graph is either in S or is adjacent to some vertex in S.

Since the IBM problem in the SCLT model is an optimization problem, we show the reduction using the following decision version of the problem.

> **Problem** (Influence Blocking Maximization in the SCLT model - Decision Version)**.**
>
> **Input:** graph $G = (V, E)$ with thresholds $t_v = \lfloor \delta_{\text{in}}(v)/2 \rfloor$ for each $v \in V$, a negative seed set S^-, and positive integers k and q.
> **Output:** Yes or No, depending if there exists a positive set of nodes S^+ of size k that blocks q of negative influence.

Theorem 1. *The IBM is NP-complete under the SCLT model.*

Proof. Given a graph G representing an instance of the DS problem, we construct a corresponding instance for the IBM problem as follows. We start by creating a directed graph G' derived from G. For each vertex v_i in G, we create a gadget structure. Each gadget consists of a *top node* that belongs to the negative seed set and two layers of *intermediate vertices*, the first contains $2|V|$ nodes, and the second contains $|V|$ nodes. The second layer connects to a *bottom node* corresponding to vertex v_i from the original graph G. The bottom nodes form a graph similar to G except that for every edge in G we create a *connection layer* in G' with $|V|$ bidirectional edges. Therefore, the total size of the constructed graph G' is $O(|V|^2 + |E| \cdot |V|)$. Figure 4.1 illustrates this construction.

We now continue with our proof, demonstrating Lemmas 2, 3, and 4.

Lemma 2. *Let v be a node in either one of the intermediate layers or the connection layer. Adding v to S^+ will only block negative influence at v itself.*

Proof. Vertices in the first intermediate layer can only directly influence vertices in the second intermediate layer. Similarly, vertices in the second intermediate layer and the connection layer can only directly influence bottom nodes. However, nodes in the second intermediate layer and the bottom nodes will not be influenced by positive seeds from the layers above. This is because their thresholds are $\frac{2|V|}{2} = |V|$, while only $k < |V|$ nodes can be selected as seeds. Therefore, including v in S^+ will only block negative influence at v itself. □

Lemma 3. *For a given k, and graphs G and G', if there exists a positive seed set S^+ of size at most k in G' that achieves a reduction of negative influence of at least $|E| \cdot |V| + |V|$ vertices, then there exists a Dominating Set (DS) of size at most k in G.*

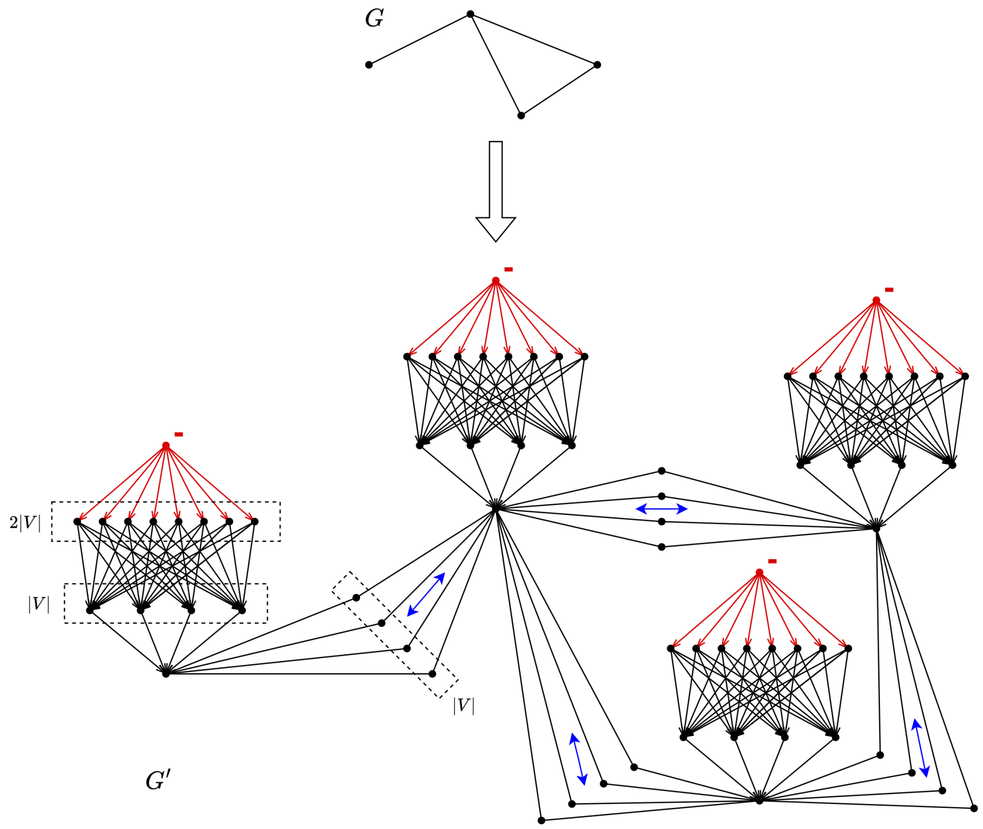

Figure 4.1: Construction of graph G'.

Proof. By Lemma 2, we can assume without loss of generality that the vertices selected for S^+ must be the bottom nodes. Note that we can only block $|E| \cdot |V| + |V|$ nodes if we influence all the bottom and connection layer nodes. The negative cascade can reach a bottom node in three time steps and a connection layer node in four time steps through the gadget structure. In comparison, the positive cascade through its seeds can reach another bottom node in two time steps and a neighboring connection layer in one time step. There might also exist edges between two bottom nodes that do not belong to S^+. In this case, the positive cascade will reach the connection layer in three time steps, which is still faster then the negative cascade. Therefore, for the positive cascade to reach all the bottom and connection layer nodes before the negative cascade, every bottom node must be connected to at least one node in S^+. Consequently, S^+ forms a dominating set of size k in the original graph G.

□

Lemma 4. *For a given k and graphs G and G', if there exists a Dominating Set (DS) of size at most k in G, then there exists a positive seed set S^+ of size at most k in G' that achieves a reduction of negative influence of at least $|E| \cdot |V| + |V|$ vertices.*

Proof. By selecting the same vertices from the dominating set as positive seeds in G', we can positively influence all the $|V|$ bottom nodes. Furthermore, as demonstrated in 3, these positive seeds cover all the connection layer nodes before the negative cascade can influence them, which accounts for $|E| \cdot |V|$ nodes. This leads to a reduction of negative influence on a total of $|E| \cdot |V| + |V|$ vertices. □

Therefore, using Lemma 3 and Lemma 4 we can conclude that by verifying whether G' has a positive seed set of size k that results in a negative influence reduction of $|E| \cdot |V| + |V|$, we can determine if the original graph G has a dominating set of size k. □

5 INTEGER LINEAR PROGRAMMING FORMULATION

This section presents an integer linear programming formulation to solve the IBM Problem in the SCLT diffusion model. The key idea of the formulation is to use the concept of time as a step-by-step way of simulating the diffusion process. To achieve this, we introduce the variable a_v. This variable denotes the elapsed time since the activation of node v. This ensures that the activation time of the source node is always greater than that of the target node when one node activates another. Thus, a node v can only influence u if $a_v > a_u$, otherwise node v would not have been activated yet. By managing activation times in this manner, we can effectively handle cascades and prevent the occurrence of cycles. Next, we present the ILP formulation.

Input Parameters

G		the input graph $G = (V, E)$
S_v^-	$v \in V$	node v belongs to S^-
t_v	$v \in V$	threshold of node v
k		the number of nodes to be selected for S^+

Variables

S_v^+	$v \in V$	node v belongs to S^+
x_v^-	$v \in V$	node v is negatively activated
x_v^+	$v \in V$	node v is positively activated
$y_{u,v}^-$	$(u, v) \in E$	u exerts negative influence over v
$y_{u,v}^+$	$(u, v) \in E$	u exerts positive influence over v
a_v	$v \in V$	elapsed time since the activation of node v

Objective Function

$$\min \quad \sum_{v \in V} x_v^- \tag{5.1}$$

Constraints

$$x_v^- \geq 1 - \frac{t_v}{\sum\limits_{u \in \mathrm{in}(v)} y_{u,v}^-} \qquad \forall v \in V \tag{5.2}$$

$$x_v^+ - S_v^+ < \frac{\sum\limits_{u \in \mathrm{in}(v)} y_{u,v}^+}{t_v} \qquad \forall v \in V \tag{5.3}$$

$$a_u + n(1 - y_{u,v}^- - y_{u,v}^+) > a_v \qquad \forall (u,v) \in E \tag{5.4}$$

$$y_{u,v}^- \leq x_u^- \qquad \forall (u,v) \in E \tag{5.5}$$

$$y_{u,v}^+ \leq x_u^+ \qquad \forall (u,v) \in E \tag{5.6}$$

$$\sum_{v \in V} S_v^+ = k \tag{5.7}$$

$$S_v^- \leq x_v^- \qquad \forall v \in V \tag{5.8}$$

$$S_v^+ \leq x_v^+ \qquad \forall v \in V \tag{5.9}$$

$$a_v + S_v^+ + S_v^- \geq 1 \qquad \forall v \in V \tag{5.10}$$

$$y_{u,v}^+ + y_{u,v}^- = 1 \qquad \forall (u,v) \in E \tag{5.11}$$

$$x_v^+ + x_v^- = 1 \qquad \forall v \in V \tag{5.12}$$

$$x_v^+, x_v^- = \{0,1\} \qquad \forall v \in V \tag{5.13}$$

$$y_{u,v}^+, y_{u,v}^- = \{0,1\} \qquad \forall (u,v) \in E \tag{5.14}$$

$$S_v^+, S_v^- = \{0,1\} \qquad \forall v \in V \tag{5.15}$$

$$a_v = \{0,1,2,\ldots,n\} \qquad \forall v \in V \tag{5.16}$$

We presented Constraint 5.2 as a non-linear constraint due to its more intuitive interpretation. However, it can be linearized by employing the "big M" method to model conditional constraints, resulting in the following inequality

$$M \cdot x_v^- \geq \sum_{u \in \mathrm{in}(v)} y_{u,v}^- - t_v, \tag{5.17}$$

where M is a sufficiently large constant that ensures the inequality holds when $x_v^- = 1$. In this formulation, if $\sum_{u \in \mathrm{in}(v)} y_{u,v}^- > t_v$, then x_v^- is set to 1 and the constant M guarantees that the constraint is satisfied. Conversely, if $\sum_{u \in \mathrm{in}(v)} y_{u,v}^- < t_v$, the right-hand side of the Equation becomes negative, resulting in x_v^- being set to 0 due to the minimization objective.

The correctness of the presented formulation is not obvious, thus Theorem 6 presents a proof of the correctness. Before presenting Theorem 6, we need a technical lemma, presented in Lemma 5, which is used later in the proof of Theorem 6.

Lemma 5. *If $\sum_{u \in in(v)} y^+_{u,v} > t_v$, then the ILP formulation sets $x^+_v = 1$.*

Proof. The SCLT Model sets the threshold to be half the number of the incoming edges. So,

$$\sum_{u \in \text{in}(v)} y^+_{u,v} > t_v = \frac{\delta_{\text{in}}(v)}{2}. \tag{5.18}$$

Furthermore, as all nodes are expected to be activated,

$$\sum_{u \in \text{in}(v)} y^+_{u,v} + \sum_{u \in \text{in}(v)} y^-_{u,v} \leq \delta_{\text{in}}(v).$$

Rearranging,

$$\sum_{u \in \text{in}(v)} y^-_{u,v} \leq \delta_{\text{in}}(v) - \sum_{u \in \text{in}(v)} y^+_{u,v}.$$

Combining with Inequality (5.18),

$$\sum_{u \in \text{in}(v)} y^-_{u,v} \leq \delta_{\text{in}}(v) - \sum_{u \in \text{in}(v)} y^+_{u,v} < \delta_{\text{in}}(v) - \frac{\delta_{\text{in}}(v)}{2} = \frac{\delta_{\text{in}}(v)}{2}$$

i.e., $\sum_{u \in \text{in}(v)} y^-_{u,v} = \delta_{\text{in}}(v)/2$. So, the right side of Constraint (5.3) becomes 0, meaning that x^-_v can be either 0 or 1. However, the objective function minimizes the negative nodes, so $x^-_v = 0$. By Constraint (5.12), we conclude that $x^+_v = 1$. □

Theorem 6. *The ILP formulation presented correctly models the IBM Problem.*

Proof. Constraint (5.2) states that a node is negatively activated when the number of incoming edges from negative neighbors exceeds its threshold. More specifically, when $\sum_{u \in \text{in}(v)} y^-_{u,v} > t_v$, the right-hand side of the inequality becomes a strictly greater than 0. Since x^-_v is a binary variable, it must be set to 1 to satisfy the constraint, indicating the negative activation of node v. So, this constraint ensures that the activation condition is met when there is a sufficient number of negative neighbors contributing to the activation of node v. Constraint (5.3) specifies that for a node to be positively activated, it must meet either one of two conditions. Firstly, the number of incoming edges from its positive neighbors must exceed its threshold. Secondly, it must be included in the positive seed set. If neither of these conditions is met, the node is not positively activated. Note that, when $\sum_{u \in \text{in}(v)} y^+_{u,v} > t_v$, node v should be activated, i.e. $x^+_v = 1$. However, by the formulation, the right-hand side of the inequality becomes greater than 1, allowing x^+_v to be either 0 or 1. That will not be a problem as, according to Lemma 5, x^+_v will be set to 1. Constraint (5.4) stipulates that if node v is activated by node u, then $a_v < a_u$. This ensures that the dissemination cascades remain acyclic, as a node cannot be activated by another node with a lower activation time. Constraints (5.5) and (5.6) ensure that a node can only be activated if its source node is also activated. By Constraint (5.7), k nodes should be selected to be in S^+. Constraints (5.8) and (5.9) require seed nodes to be activated. Constraint (5.10) determines that if a node is not in any seed set, its activation time has to be greater than 1. Constraint (5.11) states that an edge can only belong to either the negative propagation or the positive propagation, but not both. Constraint (5.12) says that a node can only be negative or positive, but not both. Constraints (5.13), (5.14), and (5.15) guarantee that x^-_v, x^+_v, $y^-_{u,v}$, $y^+_{u,v}$, S^-_v and S^+_v are binary variables. Constraint (5.16) determines that a_v is an integer variable not greater than n. □

6 COMPUTATIONAL EXPERIMENTS

In this section, we describe our experiments and discuss the obtained results.

6.1 RUNTIME ENVIRONMENT

Our computational experiments were run in an Intel Core i7-8565U 1.80GHz with 12GB of RAM, using Gurobi Optimizer 9.1.2 as the underlying LP-solver. To generate the graphs used in the experiments, we employed the NetworkX library (version 2.5.1) (Hagberg et al., 2008) in Python (version 2.7.17).

6.2 METRICS

We performed experiments to evaluate the proposed formulation. The metrics we consider are the *maximum and average empirical integrality gap* and the *running time to obtain an integer solution* using a generic branch-and-bound technique already provided by the solver.

The *integrality gap* for a minimization problem is defined as the maximum ratio between the solution value of the integer program and of its relaxation. The integrality gap is an interesting metric as it suggests how tight the relaxed formulation is compared to the convex hull of the integer solutions. Also, in some cases, the integrality gap translates into the approximation ratio of an approximation algorithm (e.g. when using variable rounding) (Young, 1995; Raghavan and Tompson, 1987). The integrality gap must be obtained analytically, but this is not always possible or simple, so we perform an empirical analysis of it. Furthermore, the integrality gap is a worst case definition, but “typical” cases may better reflect reality. This justifies our empirical *average* analysis of the integrality gap.

The running time when using a generic branch-and-bound technique serves as a baseline for other solutions, as it is an upper bound that must not be exceeded. We note that several integer linear programming solution methods can be used on our formulation, but such methods are not within the scope of this work, our focus is on the quality of the formulation itself.

6.3 INSTANCES

We use power-law graphs obtained by the Barabási-Albert model (Barabási and Albert, 1999) generated by NetworkX library as input for the Gurobi solver. Since graphs representing real networks (complex networks) are sparse, then the graphs were created with average degree 5. Each edge in the graphs was randomly assigned a direction. Additionally, each node had a 10% chance of being selected as part of the set S^-. The parameter k, determining the size of S^+, was set to half the size of S^-. The idea of using synthetic graphs in the experiments allows us to compute the average of the proposed metrics, in addition to making it possible to measure the scalability of our formulation.

6.4 RESULTS AND DISCUSSION

We solved the ILP and compared its results with the linear relaxation, where the variables in the model are allowed to take continuous values. The linear relaxation is useful to evaluate the quality of the ILP formulation, either through the execution time or the integrality gap. More specifically, we compute the average and maximum integrality gap, as well as the average time required for each model to solve the problem. For each graph size, we perform 10 executions, and for each execution, a new graph is generated. In order to obtain more reliable results, we disabled the preprocessing, cutting planes and heuristics in Gurobi, relying only on the default branch-and-bound scheme, which is the standard technique for Gurobi's integer programming solution.

The results presented in Table 6.1 demonstrate that the ILP consistently produces solutions that are similar to the relaxation. The average gap remained consistently close to 1 for all instance sizes, never exceeding 1.1. Furthermore, on average, the maximum gap is only 7% larger than the corresponding average gap, indicating that the ILP formulation performs well and maintains a relatively small variation in the quality of solutions across different instances.

In Figure 6.1 , we plot the values of Table 6.1 related to the execution time in order to get a better view, showing a comparison of the execution times between the ILP and the relaxed model. It suggests an exponential growth pattern as the graph size increases. On average, the relaxed model was approximately 79% faster than the ILP model. However, this difference tends to decrease for larger instances.

Table 6.1: ILP and linear relaxation performance.

	Average execution time (s)		Integrality gap	
Number of vertices	ILP	Linear relaxation	Maximum	Average
50	0.0501238	0.0335631	1.24138	1.02413
75	0.0577051	0.0394926	1.31506	1.07191
100	0.0898999	0.0523067	1.07692	1.0224
125	0.133355	0.0614975	1.10344	1.03172
150	0.248824	0.0811536	1.09164	1.05086
200	0.262325	0.0996094	1.0844	1.03416
400	0.316307	0.221297	1.06194	1.01856
800	1.07569	0.517779	1.0909	1.04631
1600	2.82034	2.17428	1.07035	1.0434
3200	5.85123	5.83748	1.05813	1.04229
6400	19.3689	18.7263	1.04426	1.03945
12800	163.72	134.685	1.05726	1.04090

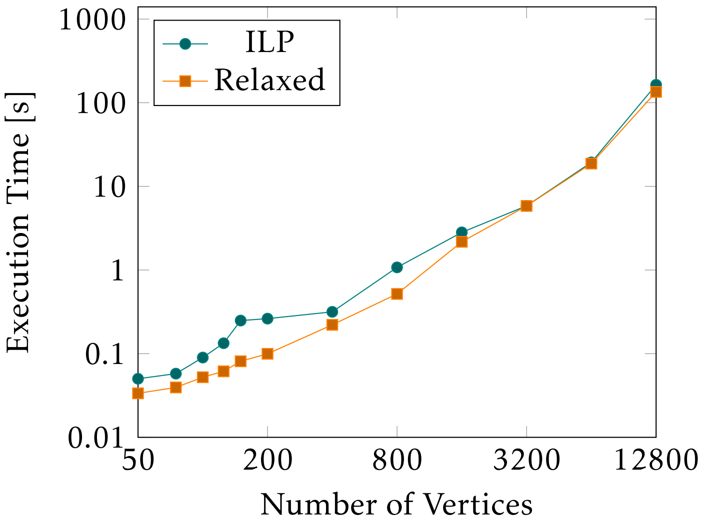

Figure 6.1: ILP and linear relaxation average execution time

7 CONCLUSION

In this work, we present a version of the Influence Blocking Maximization problem formulated under the SCLT model. We propose an integer linear programming formulation for this problem and establish its NP-hardness. Our experimental results validate the effectiveness of our approach across diverse instances, ranging from small to large scales. Importantly, our method consistently achieves results closely aligned with those obtained from its relaxed version, demonstrating both its robustness and practicality.

There is no trivial or direct way to obtain an integer linear programming formulation for this problem, and we believe that this can serve as a foundational point for the study of problems related to misinformation from the perspective of mathematical programming. This naturally motivates several future directions, such as exploring alternative formulations and variants for this problem and leveraging advanced ILP techniques to improve scalability and efficiency. Furthermore, this research provides a methodological framework adaptable to other dissemination problems involving competing influences in different domains, such as public health and politics.

In summary, our ILP formulation for the Influence Blocking Maximization problem represents a significant advancement in the field of mathematical programming applied to dissemination problems and contributes to the efforts to combat misinformation through innovative mathematical techniques. The results of our experiments pave the way for future studies aimed at refining and extending this work. The outcomes of this research were published in the Proceedings of the Brazilian Symposium on Operations Research (SBPO) 2023 (Boneti et al., 2023).

REFERENCES

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.

Boneti, G. C., Vignatti, A., and De Melo, R. S. (2023). Maximizing influence blocking with competing cascades using integer linear programming. In *55° Simpósio Brasileiro de Pesquisa Operacional (SBPO 2023)*.

Chen, N. (2009). On the approximability of influence in social networks. *SIAM Journal on Discrete Mathematics*, 23(3):1400–1415.

Cinelli, M., Quattrociocchi, W., Galeazzi, A., Valensise, C. M., Brugnoli, E., Schmidt, A. L., Zola, P., Zollo, F., and Scala, A. (2020). The covid-19 social media infodemic. *ArXiv*, abs/2003.05004.

de Melo, R. S. (2021). *Exact Algorithms For Influence Propagation In Complex Networks*. PhD thesis, Department of Computer Science, Federal University of Paraná.

Fischetti, M., Kahr, M., Leitner, M., Monaci, M., and Ruthmair, M. (2018). Least cost influence propagation in (social) networks. *Mathematical Programming*, 170(1):293–325.

Garey, M. R. and Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman, first edition edition.

Ghayour, Baghbani, F., Asadpour, M., and Faili, H. (2019). Integer linear programming for influence maximization. *Iranian Journal of Science and Technology, Tran. of Elec. Eng.*, 43(3):627–634.

Goyal, A., Lu, W., and Lakshmanan, L. V. (2011). Celf++: Optimizing the greedy algorithm for influence maximization in social networks. In *Proc. of the 20th Int. Conf. Companion on WWW*, page 47–48.

Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

He, X., Song, G., Chen, W., and Jiang, Q. (2012). Influence blocking maximization in social networks under the competitive linear threshold model. In *Proc. of the 2012 SIAM Int. Conf. on Data Mining (SDM)*, pages 463–474.

Kempe, D., Kleinberg, J., and Tardos, E. (2003). Maximizing the spread of influence through a social network. In *Proc. of the 9th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, page 137–146.

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., and Glance, N. (2007). Cost-effective outbreak detection in networks. In *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, page 420–429.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Raghavan, P. and Tompson, C. D. (1987). Randomized rounding: A technique for provably good algorithms and algorithmic proofs. *Combinatorica*, 7(4):365–374.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Young, N. E. (1995). Randomized rounding without solving the linear program. In *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 170–178. Society for Industrial and Applied Mathematics.