IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 4, 2020, accepted March 26, 2020, date of publication April 21, 2020, date of current version May 11, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2989140

Multi-Topic Misinformation Blocking With Budget Constraint on Online Social Networks

DUNG V. PHAM[1], GIANG L. NGUYEN[1], TU N. NGUYEN[2], (Senior Member, IEEE), CANH V. PHAM[3], AND ANH V. NGUYEN[1]

[1]Institute of Information Technology, Vietnam Academy of Science and Technology (VAST), Hanoi 100000, Vietnam
[2]Department of Computer Science, Purdue University Fort Wayne, Fort Wayne, IN 46805, USA
[3]ORLab, Faculty of Computer Science, Phenikaa University, Hanoi 100000, Vietnam

Corresponding authors: Dung V. Pham (pvdungc500@gmail.com) and Anh V. Nguyen (anhnv@ioit.ac.vn)

This work was supported by the Institute of Information Technology, Vietnam Academy of Science and Technology, under Project CS20.02.

ABSTRACT Along with the development of Information Technology, Online Social Networks (OSN) are constantly developing and have become popular media in the world. Besides communication enhancement benefits, OSN have such limitations on rapid spread of false information as rumors, fake news, and contradictory news. False information spread is collectively referred to as misinformation which has significant on social communities. The more sources and topics of misinformation are, the greater the number of users are affected. Therefore, it is necessary to prevent the spread of misinformation with multiple topics within a given period of time. In this paper, we propose a Multiple Topics Linear Threshold model for misinformation diffusion, and define a misinformation blocking problem based on this model that takes account of multiple topics and budget constraint. The problem is to find a set of nodes that minimizes the impact of misinformation at an allowed cost when blocking them from the network. We prove that the problem is NP-hard and the time complexity of the objective function calculation is #*P*-hard. We also prove that the objective function is monotone and submodular. We propose an approximation algorithm with approximation ratio $(1 - 1/\sqrt{e})$ based on these attributes. For large networks, we propose an extended algorithm by using a tree data structure for quickly updating and calculating the objective function. Experiments conducted on real-world datasets show efficiency and effectiveness of our proposed algorithms in comparison with other state-of-the-art algorithms.

INDEX TERMS Information diffusion, misinformation blocking, optimization, social networks.

I. INTRODUCTION

Online social networks have become one of the most efficient communication channels over the last two decades with very high socio-economic impacts. A great deal of recent research has focused on tasks of social network analysis including network modelling, network annotation, community detection, link prediction, and information diffusion. Modelling information diffusion is a key social network analysis with many useful real-world applications. For example, it can be used for early prediction of important social events, for improving recommendation performance of products, and for maximizing advertising effects to users.

In OSN, information can spread very quickly through network connections. Topics discussed and diffused on OSN can be everything from political comments, business marketing, personal concerns, to entertainment gossip. Besides many positive benefits, OSN can also bring risks to users by spreading fake news and wrong information [1], [2]. Being interested in mitigating the misinformation risks, in this paper we study the problem of modelling misinformation diffusion and propose effective methods to detect misinformation sources and limit its spread.

There have been previous studies for minimizing the impacts of misinformation diffusion in OSN [3]–[6]. A commonly used method in these studies is to disable users and connections that are considered to have major roles in spreading misinformation [7]. Finding users and connections to be disabled is addressed by solving a combination optimization

The associate editor coordinating the review of this manuscript and approving it for publication was Sherali Zeadally.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

problem. Most of these studies, however, consider only a single source of misinformation belonging to only one topic. In this paper we consider a more realistic scenario where multi-topic misinformation can reach and affect users at the same time. This problem setting poses significant challenges. First, impacts of multi-topic misinformation are proved heterogeneous [8], [9] and outcomes of the model must be re-defined. Second, when a node can adopt multiple topics, it is shown that the overall influence function that counts activated nodes is no longer submodular which is a key property to devise good approximation of the optimal solution.

We develop a new model of misinformation diffusion blocking with multiple topics and budget constraint. An important characteristic of the model is that a node in the network can be activated multiple times by multiple topics. Defining an objective function which is the influence function is defined based on this setting and is proved having monotone and submodular properties. Effective approximation methods for minimizing misinformation spread are then proposed from these monotone and submodular properties of the objective function.

The features and main contributions of this paper are as follows:

- A Multiple Topics Linear Threshold (MT-LT) model is developed by extending the Linear Threshold model [10]. Multi-topic misinformation diffusion is modelled based on different degrees of influence and activation thresholds for each topic. Then, a misinformation blocking problem, also called the Multiple Topics and Budget Constraint (MMTB) problem, is formulated by the MT-LT setting.
- We show that the MMTB problem is NP-hard and the calculation of the objective function is #P-hard. We also show that the objective function is monotone and submodular.
- Based on the monotone and submodular properties of the objective function, we propose efficient and effective algorithms for solving the MMTB problem. The first algorithm, called IGA, is an approximation greedy algorithm with approximation ratio $(1 - 1/\sqrt{e})$. The second algorithm, called GEA, is capable of running on large OSN by using a tree data structure for quickly updating and calculating the objective function.

Proposed algorithms are tested on real-world datasets including Gnutella, NetHepP and Epinions. Experimental results show that our proposed algorithms outperform other algorithms in terms of both efficiency and scalability. In particular, IGA is more effective in preventing the spread of misinformation by blocking super-influencing nodes, and GEA can be applied to medium and large networks.

Organization: The structure of the paper is organized as follows. Section I introduces an overview of the proposed work. Section II presents related works. Section III introduces the network model and research problem. Section IV presents our proposed algorithms and section V provides experimental results on some selected datasets. Section VI concludes the paper.

II. RELATED WORKS

Kempe *et al.* [10] formulated stochastic discrete optimization problem under the Independent Cascade Model (IC model), and Linear Threshold Model (LT model). Kempe's research is inspired by Sebastos and Richardson's research on information spreading using data mining techniques [11]. The problem in [10] is formulated as follows: given a network, a diffusion model, a set of influence weights for edges, a random threshold function, and a budget; selecting nodes so that the final number of infected nodes is maximized. For the problem they formulated, Kempe *et al.* proposed a greedy algorithm with an approximation guarantee of $(1 - 1/e)$. Later, many studies on information diffusion and misinformation spreading prevention problems on online social networks have been undertaken [12]–[14]. The authors in [15] studied the problem of eliminating k-edge sets so that influence of the S-sourceset is minimal and introduced an algorithm for approximation $(1 - 1/e - \epsilon)$, wherein e and $\epsilon \in [0, 1]$ to solve the problem. From the epidemiological perspective, some authors injected immunization vaccines into sets of vertices to be immune to bad information [4], [5], [16]. The authors in [17], [18] studied the DAVA probslem (Data-Aware Vaccination) with a request to inject the vaccine into the k-vertices of the user set. In [19], the authors extended the DAVA problem by adding time to spreading the disease.

On the other hand, many researches followed an approach of spreading good information to prevent impact of bad information called information purification method [3], [20], [21]. The authors in [22] proposed the MCIC (Multi - Campaign Independent Cascade) information diffusion model that allows multiple sources of information to be spread simultaneously on the same network. For the same purpose, in [23] the authors studied to prevent influence of misinformation on the linear competition model. In addition, the authors in [6] studied the TIB (Temporal Influence Blocking) problem to limit misinformation by time delay. The authors in [24] studied the new β_T^I problem with a goal of selecting the smallest seed set to start spreading good information to eliminate bad information.

Recently, the author in [25] studied misinformation containment with multiple cascades. In [26] the authors investigated rumor blocking within a given community. In [27] the authors proposed a scalable algorithm which guarantees approximation ratio of $(1 - 1/e - \epsilon)$ for epidemic blocking problem by edges and nodes blocking. In [28] the authors studied influence blocking which considers location of competitors. The authors in [29], [30] proposed a method for misinformation prevention by eliminating nodes in multiple contexts. Furthermore, several studies have focused on identifying and detecting misinformation which is an important step for issues that prevent misinformation. The authors in [31], [32] relied on structure and language characteristics to identify false information. Some studies used data

mining and machine learning methods to detect misinformation by user behavior analysis such as shares, comments, and likes [33]–[35].

III. MODEL AND PROBLEM FORMULATION

The IC model and the LT model are two of the most widely used models in the research of information diffusion problem on online social networks. In the IC model, an active node u may attempt to activate a neighboring inactive node v only once with successful probability $p(u, v)$. IC model can be seen as a sender-central model. In the LT model, every node contributes to activation of their neighbors. So, LT model can be considered as a receiver-central model. With reference to the problem of preventing spreading misinformation, the LT model, having more advantages, is more well-suited than the IC model. The collective contribution of active nodes in activating their neighbors in the LT model can be seen as herding effect, which is very close to mechanism of spreading false rumors where the decision is more likely to be made by mimicking others' decision. In this section, we formulate a Multiple Topics Linear Threshold (MT-LT) model by extending the LT model. This MT-LT model considers multi-topic misinformation diffusion with budget constraint. Next, we present the traditional LT model. All the symbols and notations used in the paper are given in Table 1.

TABLE 1. Symbols and notations.

Notional	Description
n, m	The number of nodes and the number of edges in graph G.
$N_{in}(v)$, $N_{out}(v)$	The sets of incoming and outgoing neighbor nodes of v
S	The set of nodes as the source of misinformation
A	The set of nodes which is blocked from the network
p_v^i	The effect of node v on its neighboring nodes by topic i
γ_v^i	The activation threshold of node v on topic i
$\mathcal{D}_{LT}(G_i, S_i)$	Influence function for source set S_i in graph G_i under LT model.
$\mathcal{D}(G, S)$	Influence function for source set S in graph G under MT-LT model.
$G \odot A$	The graph after blocking a set of nodes A.
$P(G, v)$	The set of all simple paths starting node v in G
$\sigma(G, S, A)$	Influence reduction function for source set S after blocking in graph G under MT-LT model (objective function).

A. INFORMATION DIFFUSION MODEL

1) LINEAR THRESHOLD MODEL

In the LT model, an online social network is represented by a graph $G = (V, E, w)$ in which V is a node set, E is a directed edge set, $|V| = n$, $|E| = m$ and $N_{in}(v)$, $N_{out}(v)$ are the set of incoming neighbor nodes, outgoing neighbor nodes of node v, respectively. Each edge $(u, v) \in E$ is assigned with a weight $w(u, v) \in [0, 1]$ representing the influence of node u on node v, if $w(u, v) \notin E$ then $w(u, v) = 0$. Weights are distributed such that the sum of weights of neighboring nodes to a node v satisfies the following condition: $\sum_{u \in N_{in}(v)} w(u, v) \leq 1$

Suppose that $S_0 \subseteq V$ is the set of nodes which spreads misinformation and it is called the *seed set*. In LT model, each node may have one of two states: *active* and *inactive*.

Each node $v \in V$ has an activation threshold $\gamma_v \in [0, 1]$, if γ_v is large, many neighbor nodes are required to activate v, if γ_v is small, node v can be easily activated by its neighbors. In many related works, threshold values are determined randomly over the [0, 1] segment. In practice, threshold values can be learned via data mining techniques based on user actions in the past. Thus, threshold values can be viewed as an input to the model instead of assuming a random threshold function. Let $\mathcal{D}^t(G, S)$ the set of nodes activated by S at time step t in graph $G(V, E, w)$. The LT model operates in discrete time steps as follows:

- At time step $t = 0$, the set of nodes in the active state is the source of the original information diffusion S_0 (*seed set*).
- At time step $t \geq 1$, all nodes activated by S in time step $t - 1$ are still active. A node v currently not activated by S will become activated if the following condition satisfies:

$$\sum_{u \in N_{in}(v) \cap \mathcal{D}^{t-1}(G,S)} w(u, v) \geq \gamma_v.$$

- The diffusion process ends when no node is activated in the next steps.

2) MULTIPLE TOPICS LINEAR THRESHOLD

The LT model considers diffusion of a single topic, or single information cascade. Motivated by LT model, a more realistic scenario is studied in this paper where we assume that there are multiple existing topics being diffused. Topics may have different characteristics, such as their content and impressiveness. When there are multiple topics, we need to redefine outcomes of the model when two or more topical information reach one user at the same time. The LT model can not be applied directly to solve the problem of multi-topic information diffusion because it is hard to capture complex correlations between topical cascades.

Earlier researchers have worked on a scenario where there are more-than-one topics being diffused. When multiple topics exist, the influence maximization problem can be elusive as even not being monotone [25]. When a node can adopt multiple cascades, it is shown that the overall influence function that counts activated nodes is no longer submodular. In this paper, we deal with this problem by developing a new model of misinformation diffusion blocking with multiple topics and defining the overall influence function that counts *activated turns* instead of activated nodes.

In MT-LT model, a social network is also represented by a graph $G = (V, E, w)$ where V is the set of nodes, E is the set of edges, $|V| = n$, $|E| = m$. $N_{in}(v)$, $N_{out}(v)$ are the set of incoming neighbor nodes, outgoing neighbor nodes

of node v, respectively. Each edge $(u, v) \in E$ is assigned a weight $w(u, v) \in [0, 1]$ representing the influence of node u on node v, if $(u, v) \notin E$ then $w(u, v) = 0$. Weights are distributed such that the sum of the weights of nodes u to node v satisfies $\sum_{u \in N_{in}(v)} w(u, v) \leq 1$.

Suppose that there are q misinformation topics such as Economics, Politics, Sports, and so on, and the set of misinformation-spreading nodes is $S = \{S_1, S_2, \ldots, S_q\}$. S_i contains nodes spreading information of topic i (referred to as source nodes). We can assume that the social network administrator knows where the source of misinformation is. The set of source nodes spreading misinformation on q topics is $S = \bigcup_{i=1}^{q} S_i$.

Each node $v \in V$ may be activated multiple times by multiple topics. That means, node v may have multiple statuses in the set of $q + 1$ status as follows: $Q = \{inactive, active\_1, active\_2, \ldots, active\_q\}$ which shows the behavior and activity of v. If node v is *inactive* then v is not activated by any topic; if v is $active\_i$ then it has been activated by topic i. If node v has the status $active\_1, active\_2, \ldots, active\_k, 1 \leq k \leq q$ then it has been activated by k topics.

In practice, the impact weight among nodes depends on topics. For example, a spreading topic about a plague may have greater impact than a topic about a sport game to an user. Therefore, a node v is assigned with a vector of activation thresholds $\gamma_v = (\gamma_v^1, \gamma_v^2, \ldots, \gamma_v^q)$, where $\gamma_v^i \in [0, 1]$. γ_v^i represents the activation threshold of node v on topic i. Moreover, each node v is also assigned with a vector $P_v = (p_v^1, p_v^2, \ldots, p_v^q)$, where $p_v^i \in [0, 1]$ represents the effect of v on its neighboring nodes by topic i.

The process of information spread in model MT-LT occurs in separated time steps $t = 1, 2, \ldots, d$ where $d \in Z$. We consider the same allowed period for each step of information spread. It is because all neighbors of a node might not influence it simultaneously, but within a certain time window. Let $\mathcal{D}_i^t(G, S)$ be the set of nodes activated by S_i at time step t in graph G.

- At time step $t = 0$, all nodes in S_i have the status $active\_i$
- At time step $t \geq 1$, all nodes activated by S_i in time step $t-1$ are still active. A node v currently not activated by S_i will become $active\_i$ if the following condition satisfies: $\sum_{u \in N_{in}(v) \cap \mathcal{D}_i^{t-1}(G,S)} w(u, v) \cdot p_u^i \geq \gamma_v^i$.
- The spread process ends when no node is activated in the next steps.

The LT model is a special case of MA-LT model when $p_i^u = 1$, $i = 1..q$ for all $u \in V$. Let $\mathcal{D}_i(G, S)$ be the total number of nodes activated by topic i after the spreading process ends. $\mathcal{D}_i(G, S)$ is calculated by the summation of $\mathcal{D}_i^t(G, S)$ over all time steps. The total number of *activated turns* by all topics after the spreading process ends, denoted by $\mathcal{D}(G, S)$, is given by:

$$\mathcal{D}(G, S) = \sum_{i=1}^{q} \mathcal{D}_i(G, S) \tag{1}$$

In this setting a node can be activated multiple times by multiples topics not just once as in previous works. By this setting we can prove the monotone and submodular properties of the overall influence function. These properties are important because they can help to devise effective approximation algorithms.

B. PROBLEM DEFINITION

In this paper, we aim at blocking a set of nodes in the graph G so that the final number of infected turns is minimized. This similar optimization objective can be found in other works [5], [16], [36]. A blocked node cannot be infected by any other nodes, and it cannot infect any other nodes as well. To block a node v in a graph, we simply set the weights for all incoming edges to v and outgoing edges from v to zero.

Given a graph G, we denote by $G \odot A$ the graph after blocking a set of nodes A. The number of all activated turns by all topics after blocking the set of nodes A is given by $\mathcal{D}(G \odot A, S) = \sum_{i=1}^{q} \mathcal{D}_i(G \odot A, S)$. The objective here is to minimize $\mathcal{D}(G \odot A, S)$. This is equivalent to maximizing the following quantity:

$$\sigma(G, S, A) = \mathcal{D}(G, S) - \mathcal{D}(G \odot A, S) \tag{2}$$

Suppose that blocking a node v costs $c(u)$ and the total of costs cannot exceed the budget limit B. The MMTB problem can now be formulated as follows.

Definition 1: (MMTB). Given a social network represented by a weighted graph $G(V, E, w)$, a source of misinformation with q topics given by $S = \{S_1, S_2, \ldots, S_q\}$ where S_i contains source nodes of topic i. The task is to find a set of nodes A to block so that the quantity $\sigma(G, S, A)$ is maximized under the budget limit B, $c(A) = \sum_{u \in A} c(u) \leq B$.

We prove that the MMTB is NP-hard under the MT-LT setting.

Theorem 1: The MMTB is an NP-hard problem.

Proof: To prove MMTB as an NP-hard problem, we construct a derivative problem from the well-known Knapsack problem which is also NP-complete.

Knapsack problem: Given a set Q of n items, each item i has a weight w_i and a value c_i (c_i and w_i are integers) and two positive integers: W, C. The problem is to find a vector $x = (x_1, x_2, \ldots, x_n)$ so that $value(Q) = \sum_{i=1}^{n} x_i c_i \geq C$ and $\sum_{i=1}^{n} x_i w_i \leq W$ are satisfied.

Let $I_1 = (Q, W)$ be an instance of the Knapsack problem, and $I_2 = (G, S, B)$ be an instance of the MMTB problem where S is the set of misinformation source nodes, B is the budget limit, we construct an reduction from I_1 to I_2 as shown in Fig. 1.

Reduction: To construct the reduction, we construct a graph $G(V, E, w)$ satisfying the MT-LT model as follows. Given the set S with a single node $S = \{s\}$. For each c_i (the value of the i-th item) we create a path of $c_i + 1$ nodes: $s \rightarrow u_{i,1} \rightarrow u_{i,2} \ldots \rightarrow u_{i,c_i}$ with the weight of each edge of the path is 1. The cost of nodes is set as follow: $c(u_{i,1}) = w_i = 1, c(u_{i,j>1}) = B + 1$. The number of topics is $q = 1$ and the budget is $B = W$. Set $B = W$ and $K = C$.

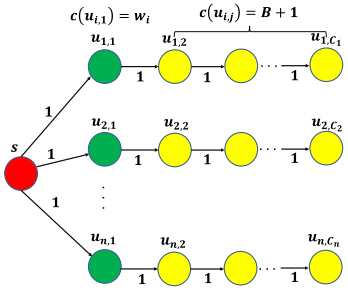


FIGURE 1. Building a reduction from Knapsack to MMTB.

We prove that I_1 has the solution $x = (x_1, x_2, \ldots, x_n)$ if and only if I_2 has the corresponding solution $A = \{u_{i,1}|x_{i=1}\}$ such that $\sigma(G, S, A) \geq K$ and vice versa.

($\rightarrow$)Assume that $x = \{x_1, x_2, \ldots, x_n\}$ is the solution of I_1, on I_2 we select $A = \{u_{i,1}|x_i = 1\}$. We have $c(A) = \sum_{i|x_{i=1}} x_i w_i = \sum_{i=1}^{n} w_i \leq W = B$. According to the MT-LT model, when blocking the node $u_{i,1}$ all sub-sequence nodes on the path $u_{i,2} \ldots u_{i,c_i}$ are affected. Thus, $\sigma(G, S, A) = \sum_{i|x_{i=1}} x_i c_i = \sum_{i=1}^{n} c_i \geq C = K$, then A is the answer of I_2.

($\leftarrow$) By contrast, if A is a solution of I_2 then A cannot contain the node $u_{i,j\geq 2}$, because the cost of blocking the nodes will exceed B. On I_1 we select a vector $x = \{x_1, x_2, \ldots, x_n\}$ on a condition that: $x_i = 1$ if $u_{i,1} \in A$ or $x_i = 0$, if $u_{i,1} \notin A$. We have $c(A) = \sum_{i|u_{i,1}\in A} w_i = \sum_{i=1}^{n} x_i w_i \leq B$ and $\sigma(G, S, A) = \sum_{i|u_{i,1}\in A} c_i = \sum_{i=1}^{n} x_i c_i \geq K = C$, that means x is an answer of I_1. In other words, if we can find the optimal solution of the MMTB problem, we can find the optimal solution of Knapsack problem. Thus MMTB problem is NP-hard. ∎

We now prove that the problem of calculating the objective function in formula 2 is #P-hard.

Theorem 2: The problem of calculating function $\sigma(\cdot)$ is #*P*-hard in MT-LT even if the set of A has only one node.

Proof: We prove that the calculation of the objective function is #*P*-hard even for the case the set S has only one node. Let $P(G, s)$ be the set of all simple paths of G starting from s (simple paths are paths that visit each node just one time), $P(G \odot A, s)$ be the set of all simple paths of G starting from s when A is blocked. We have that $\sigma(G, S, A) = \mathcal{D}(G, S) - \mathcal{D}(G \odot A, S)$ is exactly the number of nodes in $P(G, s)$ minus the number of nodes in $P(G \odot A, s)$. If we can calculate the number of nodes in $P(G, s)$ then we can also count the number of simple paths in $P(G, s)$. Counting all such simple paths is exactly the *s-t paths problem* which is proved #*P*-hard by Valiant [37]. Therefore, our problem is also #*P*-hard. ∎

IV. PROPOSED ALGORITHMS

In this section we propose two algorithms to solve the MMTB problem. Both algorithms are based on greedy algorithm approach. The first algorithm called Improved Greedy Algorithm (IGA) is based on the ratio between the increase degree of target function and the cost of blocking the node ensuring approximation ratio $(1 - 1/\sqrt{e})$. The second algorithm called Greedy Extension Algorithm (GEA) is based on the idea of quickly updating the target function and the approximate average denominator calculating method.

A. IMPROVED GREEDY ALGORITHM-IGA

First, we show that the target function $\sigma(G, S, A)$ is monotone and submodular. Based on these features, and by adopting the greedy strategy proposed in [19] we are able to obtain an algorithm with approximation ratio $(1-1/\sqrt{e})$. The proposed algorithm is called IGA (Improved Greedy algorithm).

From each original graph $G = (V, E, w)$ under MT-LT model, we construct q graphs: $G_1, G_2, \ldots, G_q$, $G_i = (V_i, E_i, w_i)$, with $w_i(u, v) = w(u, v).p_u^i$. We show that the total number of activated turns on graph G on the MT-LT model with source S is equal to the number of nodes activated on graph G_i on the LT model with source S_i, for any $i = 1, 2, \ldots, q$. This result is proved in the following lemma:

Lemma 1: Denoted $\mathcal{D}_{LT}(G_i, S_i)$ as the set of nodes activated by source S_i on graph G_i with model LT, we have $\mathcal{D}_i(G, S) = \mathcal{D}_{LT}(G_i, S_i)$. Then the number of turns activated by all topics $\mathcal{D}(G, S)$ can be calculated as follows:

$$\mathcal{D}(G, S) = \sum_{i=1}^{q} \mathcal{D}_i(G, S) = \sum_{i=1}^{q} \mathcal{D}_{LT}(G_i, S_i) \tag{3}$$

Proof: Because $p_u^i \leq 1$, for each node $u \in G_i$ we have: $\sum_{u\in N_{in}(v)} w_i(u, v) \cdot p_u^i \leq \sum_{u\in N_{in}(v)} w_i(u, v) \leq 1$
This condition satisfies the LT model. Let $\mathcal{D}_{LT}(G_i, S_i)$ be the influence function for the source set S_i in graph G_i under LT model, we obtain $\mathcal{D}_i(G, S) = \mathcal{D}_{LT}(G_i, S_i)$. ∎

Lemma 2: For graph G_i, function $\mathcal{D}_{LT}(G_i \odot A, S_i)$ is monotone and supermodular.

$$\begin{aligned}&\mathcal{D}_{LT}(G_i \odot (A \cup \{v\}), S_i) - \mathcal{D}_{LT}(G_i \odot A, S_i)\\ &\quad \leq \mathcal{D}_{LT}(G_i \odot (T \cup \{v\}), S_i) - \mathcal{D}_{LT}(G_i \odot T, S_i)\\ &\quad \forall A \subseteq T \subset V, v \in T \setminus A.\end{aligned}$$

Proof: Let $E(A)$ be the set of edges which have at lest a node in node set A. We have: $\mathcal{D}_{LT}(G_i \odot A, S_i) = \mathcal{D}_{LT}(G_i \odot E(A), S_i)$.

It is obvious that $\mathcal{D}(G \odot E(A), S) - \mathcal{D}(G \odot E(T), S) \geq 0$ for $A \subseteq T$. Therefore $\mathcal{D}_{LT}(G_i \odot A, S_i)$ is a monotonically increasing function.

Denote $E_{T,v} = E(T \cup \{v\}) \setminus E(T)$, $E_{A,v} = E(A \cup \{v\}) \setminus E(A)$. $E_{T,v}$ is the set of edges connecting to v but not to any node in the set T, $E_{A,v}$ is the set of edges connecting to v but not to any node in the set A. We have $E_{T,v} \subseteq E_{A,v}$ for $A \subseteq T$. We easily see that $E(A) \cup E_{T,v} \subseteq E(A + \{v\})$. Given two set of edges X, Y, $X \subseteq Y \subset E$, an edge $e \in Y \setminus X$. By Theorem 6 in [16], we have:

$$\begin{aligned}&\mathcal{D}_{LT}(G_i \odot E(X \cup \{e\}), S_i) - \mathcal{D}_{LT}(G_i \odot E(X), S_i)\\ &\quad \leq \mathcal{D}_{LT}(G_i \odot E(T \cup \{e\}), S_i) - \mathcal{D}_{LT}(G_i \odot E(T), S_i)\end{aligned}$$

Applying the above inequality we have:

$$\begin{aligned}
&\mathcal{D}_{LT}(G_i \odot A, S_i) - \mathcal{D}_{LT}(G_i \odot (A \cup \{v\}), S_i) \\
&\quad = \mathcal{D}_{LT}(G_i \odot E(A), S_i) - \mathcal{D}_{LT}(G_i \odot E(A \cup \{v\}), S_i) \\
&\quad \geq \mathcal{D}_{LT}(G_i \odot E(A), S_i) - \mathcal{D}_{LT}(G_i \odot (E(A) \cup E_{T,v}), S_i) \\
&\quad \geq \mathcal{D}_{LT}(G_i \odot E(T), S_i) - \mathcal{D}_{LT}(G_i \odot (E(T) \cup E_{T,v}), S_i) \\
&\quad = \mathcal{D}_{LT}(G_i \odot T, S_i) - \mathcal{D}_{LT}(G_i \odot (T \cup \{v\}), S_i)
\end{aligned}$$

This complete the proof. ■

Theorem 3: The function $\sigma(\cdot)$ is submodular and monotone on the MT-LT model.

Proof: From the definition of $\sigma(G, S, A)$ in Eq. (2), we have:

$$\begin{aligned}
\sigma(G, S, A) &= \mathcal{D}(G, S) - \mathcal{D}(G \odot A, S) \\
&= \sum_{i=1}^{q} \mathcal{D}_i(G_i, S_i) - \sum_{i=1}^{q} \mathcal{D}_i(G_i \odot A, S_i) \\
&= \sum_{i=1}^{q} \big(\mathcal{D}_i(G, S_i) - \mathcal{D}_i(G_i \odot A, S_i)\big) = \sum_{i=1}^{q} \sigma_i(G_i, S_i, A)
\end{aligned}$$

in which $\sigma_i(G, S_i, A) = \mathcal{D}_i(G, S_i) - \mathcal{D}_i(G \odot A, S_i)$. According to Lemma 2, $\mathcal{D}_i(G \odot A, S_i)$ is supermodular, and $D_i(G, S_i)$ is monotone and submodular. Therefore, $\sigma_i(G, S_i, A)$ is monotone and submodular function. $\sigma(G, S, A)$ is a collection of monotone and submodular functions, so it is also a monotone and submodular function. ■

Algorithm 1 Improved Greedy Algorithm (IGA)

Input: $G = (V, E, w)$, source set S, budget $B > 0$
Output: set of nodes A

$A_1 \leftarrow \emptyset$; $U \leftarrow V$;
$v_{max} = \arg\max_{v \in V,\ c(v) \leq B} \sigma(G, S, v)$;
repeat
  $u \leftarrow \arg\max_{v \in V \setminus A} \delta(v)$;
  if $c(A_1) + c(u) \leq B$ **then**
    $A \leftarrow A_1 \cup \{u\}$;
  end
until $U = \emptyset$;
If $\sigma(G, S, A_1) \geq \sigma(G, S, v_{max})$ **then** $A \leftarrow A_1$ **else** $A \leftarrow v_{max}$;
return A.

Based on the results of Theorem 3 and using the greedy strategy proposed in [38], we propose an innovative greedy algorithm called IGA that has approximation ratio $(1-1/\sqrt{e})$ (Algorithm 1). The algorithm consists of 2 phases. The first phase uses greedy strategy to find the set of nodes to block A. In each step, we choose a node v with $\delta(v)$ is the largest. $\delta(v)$ is calculated as follows:

$$\delta(v) = \frac{(\sigma(G, S, A \cup \{v\}) - \sigma(G, S, A)}{c(v)} \tag{4}$$

The process ends when the cost for blocking nodes exceeds the allowed budget B. In the second phase, a node v_{max} with the largest $\sigma(G, S, v_{max})$ and the cost for blocking v_{max} which less than B are considered. Then the final outcome of A is compared to v_{max} to obtain the best answer.

It is easy to see that in the worst case, algorithm IGA can take up to k^2 loop to re-calculate $\sigma(G, S, A)$, where k is the number of activated turns on q topics. However, calculating the exact number of activated turns is #P-hard. To solve this problem, we use Monte Carlo (MC) simulation method to estimate target function (Algorithm 2). With each set S_i, $i = 1, 2, \ldots, q$, we use MC simulation T times to simulate the process of random information spreading. Each time, the number of activated turns by topic i is calculated, then the average number per T simulation times is calculated. Finally, we get the average number of activated turns on q topics. The larger the number of simulations T is, the higher the estimation accurate is.

Algorithm 2 Algorithm to Estimate the Value of the Function $D_i(G_i, S_i)$

Input: $G_i(V_i, E_i, w_i)$, source set S
Output: $D_i(G_i, S_i)$

$count \leftarrow 0$;
for $i = 1$ *to* T **do**
  Simulating the misinformation propagation process from the source S_i on graph G_i;
  $N_i \leftarrow$ the number of nodes activated after the propagation has finished;
  $count \leftarrow count + N_i$;
end
return $count/T$.

However, because calculating $\sigma(G, S, A)$ is #P-hard, it is difficult to determine the number of simulations. In this case we perform T times of MC simulation, the time complexity of IGA is $O(TRn^2)$ where R is the time complexity of a MC simulation. It means that IGA cannot run on networks with even small size. For this reason, in the following subsection, we develop a more practical algorithm called GEA that can work on large networks.

B. GREEDY EXTENSION ALGORITHM-GEA

In this subsection we propose an expanded version of the greedy algorithm IGA, called Greedy Extension Algorithm (GEA). The algorithm GEA is based on the idea of calculating average value of denominator and fast updating the target function $\sigma(\cdot)$. To do so, a tree structure is used to estimate and update $\sigma(.)$ in each loop of the algorithm. We construct q graphs G_i, $i = 1, 2, \ldots, q$, according to MT-LT model and the result of lemma 1. Because the source set S_i could have more than one node, neighboring nodes of S_i could be infected by nodes of the same topic. In order to update target function conveniently and ensure the spreading properties of the model LT, we merge source nodes S_i on

Algorithm 3 Algorithm of Merging Vertices Merge(G_i, S_i)

Input: $G_i = (V_i, E_i, w_i)$, source set S_i
Output: G'_i, H_i

1. $G'_i \leftarrow G_i$
2. Add node H_i to G'_i;
3. **for** $x \in S_i$ **do**
4. **if** *there exits edge* (x, v) **then**
5. **if** $H_i \in G'_i$ **then**
6. Add edge (H_i, v) to G'_i;
   $w'_i(H_i, v) = w'_i(x, v).p_x^i$;
7. **else**
8. $w'_i(H_i, v) = w'_i(H_i, v) + w'_i(x, v).p_x^i$
9. **end**
10. ; Blocking (x, v) from S'_i;
11. **end**
12. **end**
13. Blocking all node S_i from G'_i;
14. Return G'_i, H_i.

graph G_i into a node H_i and obtain graph G'_i (Algorithm 3). Lemma 3 shows that two expressions before and after converting are equivalent.

Lemma 3: Algorithm 3 shows that any expression (G_i, S_i, w_i) is equivalent to the expression (G'_i, H_i, w'_i), where H_i is the unified source node of the nodes in S_i.

Proof: To prove this lemma, we prove that the function $\sigma(G, S, A)$ on the two expressions is the same. Assume that v is a node adjacent to the set S. When S has a single node u, it is obvious that the influence from S to v is $w(u, v).p_i^u = w(H_i, v)$. When the set S has k source nodes $u_1, u_2, \ldots, u_k$, the effect of k nodes on node v is: $\sum_{i=1}^{k} w(u_i, v).p_i^{u_i} = w(H_i, v)$, this satisfies the spreading property on the LT model. When S has $k+1$ source nodes, $w(H_i, v)$ is the effect after mixing k nodes. Mixing more $k+1$ source node, total effect $w(H_i, v) = w(H_i, v) + w(u_{k+1}, v)$, which is the effect from H_i to v according to Algorithm 3. According to the inductive proposition, we have the proof. ∎

For each graph G_i after source nodes being merged, we use Monte Carlo simulation to create n_i sample graphs g from G_i using the online edge model [10]. Because we can access nodes from trees with roots H_i, we only retain trees that can access to other nodes from the root node for the sample graph. This reduces a significant number of pointless sample graphs, helping to update values to get closer approximations. From n_i samples, we set $\mathcal{T}_i (i = 1..q)$ containing the root of n_i trees. For each tree $T_i \in \mathcal{T}_i, f(T_i, A_i)$ is the value of $\sigma(G_i, S_i, A_i)$ on tree T_i. We observe that $f(T_i, u) = |\{v | v \in subtree(u)\}|$ and we can compute $f(T_i, u)$ for all nodes $u \in T_i$ using the deep first search in algorithm 4. Since the limited budget B is used and a node may be present on many different trees, we apply sample average approximation to calculate $\sigma(G, S, A)$ on q

Algorithm 4 Calculate $f(T_i, u)$

Input: A tree T_i root at H_i and node $u \in V$
Output: $f(T_i, u)$

1 **if** *u is not a leaf* **then**
2 $r \leftarrow 1$;
3 **for** *v is a child of u* **do**
4 $r \leftarrow r + f(T_i, v)$;
5 **end**
6 **else**
7 $r \leftarrow 1$
8 **end**
9 **return** r.

topics as follows:

$$\sigma(G, S, A) \approx \hat{\sigma}(G, S, A) = \frac{1}{q} \sum_{i=1}^{q} \left(\frac{1}{n_i} \sum_{T_i \in \mathcal{L}_i} f(T_i, A_i) \right) \quad (5)$$

In Algorithm 5, for the first selection, we select set A_1 in the loop (from line 9 to line 22) by gradually adding node u into the set A_1 in a greedy manner, such that $\delta(u)$ reaches the maximum value (line 12). For the second selection, we select node u_{max} so that $c(u_{max}) \leq B$ and the sample average approximation $\sigma(u_{max})$ is maximized (line 23). Let A_1 be the current solution in the loop t, we estimate the objective function which will increase gradually by blocking node v according to the following equation:

$$\begin{aligned} \delta(A_1, v) &= \sigma(G, S, (A_1 \cup \{v\})) - \sigma(G, S, A_1) \\ &\approx \frac{1}{q} \sum_{i=1}^{q} \sum_{T_{H_i} \in \mathcal{L}_i} (f(T_i, H_i) - f(T_i \odot \{v\}, H_i) \end{aligned} \quad (6)$$

After selecting u into H_i, we conduct a calculations of all trees $T_i \in \mathcal{T}_i$ on all topics to block node u from trees T_i and update $f(T_i, u)$ on $T_i \in \mathcal{T}_i$, (lines 17-19) as follows:

1) if v is a descendant of u, we can block them because it is not reachable from $H_i \in T_i$
2) if v is a ancestor of $u, f(T_{H_i} \odot \{u\}, v) = f(T_i, v) - f(T_i, u)$

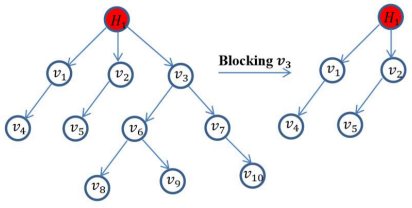


FIGURE 2. Estimating and update $f(T_i \odot \{u\}, H_i)$ by using rooted trees.

The updating process is illustrated in the Fig. 2. We can calculate $f(T_i, H_i) = 10, f(T_i, v_3) = 6$, after blocking v_3, and update $f(T_i \odot \{u\}, H_i) = 10 - 6 = 4$.
Finally, the algorithm returns a better solution in two candidate solutions u_{max} and A_1 by comparing $\hat{\sigma}(u_{max})$ and $\hat{\sigma}(A_1)$.

Algorithm 5 Greedy Extension Algorithm (GEA)

Input: Graph $G = (V, E, w)$, source S, budget $B > 0$
Output: The set of nodes A

$U \leftarrow V$;
Build $G_i = (V_i, E_i, w_i)$, $i = 1..q$ from $G = (V, E, w)$ by MT-LT model;
$(G'_i, H_i) \leftarrow Merge(G_i, S_i)$ for $i = 1 \ldots q$ according to Algorithm 3;
foreach G'_i **do**
    Generate n_i sample graphs by live-edge model [10] and create a set $\mathcal{T}_i$ contains n_i trees;
    For each $T_i \in \mathcal{T}_i$, calculate $\sigma(T_i, u)$ for all $u \in T_{H_i}$ by Algorithm 4;
end
$u_{max} \leftarrow \arg\max_{v \in V,\ c(v) \le B} \hat{\sigma}(v)$;
repeat
    $c_{min} \leftarrow \arg\min_{v \in V} c(v)$;
    If $c_{min} + c(A_1) > B$ **then** break;
    $u \leftarrow \arg\max_{v \in V} \delta(A_1, v)$ (Eq. 6);
    $U \leftarrow U \setminus \{u\}$;
    if $c(A_1) + c(u) \le B$ **then**
        $A_1 \leftarrow A_1 \cup \{u\}$;
        for $i = 1$ *to* q **do**
            foreach $T_i \in \mathcal{T}_i$ **do**
                If $u \in T_i$ **then** block node u and update $f(T_i, v), \forall v \in T_i$;
            end
        end
    end
until $U = \emptyset$;
$A \leftarrow \arg\max_{u_{max}, A_1} \{\hat{\sigma}(u_{max})), \hat{\sigma}(A_1)\}$;
return A.

Next, the time complexity of GEA is given. The time complexity of creating a set of T_i is $(n_i(n+m))$. The time complexity of calculating $f(T_i, u)$ is the same as Algorithm 4, that is $O(n)$. Each step of selecting node u with maximum value $\delta(A, u)$ needs $O(n_i n)$. Consequently, the time complexity of the algorithm GEA is $O((\sum_{i=1}^{q} n_i)(m + kn))$ where q is the number of topics, n, m are the number of nodes, the number of edges of the graph $G(V, E, w)$ respectively, and n_i is the number of trees created in the MC simulation with the topic i.

V. EXPERIMENTS RESULTS

In this section, we conduct experiments to show the efficiency of the proposed algorithms IGA and GEA. The proposed algorithms are compared with Degree and Random algorithms on the same setting of the MT-LT model.

A. EXPERIMENT SETTINGS

Datasets and parameter settings: The experiments are performed on 03 datasets, Grutela [39], Epinions [40] and NetHepPh [41], of the actual networks with size of up to tens of thousands of nodes and hundreds of thousands of edges, collected from the source [http://snap.stanford.edu/data/]. Some statistics of the datasets are provided in Table 2.

TABLE 2. Datasets.

Dataset	Nodes	Edges	Type	Avg.degree
Gnutella [44]	6K	20K	Directed	3.29
NetHepP [46]	34K	421K	Directed	12.2
Epinions [45]	75K	508K	Directed	6.7

All the algorithms are programmed in Python language. All the experiments are conducted on a computer with CPU Intel Core i7 - 8550U 1.8Ghz, RAM 8GB DDR4 2400MHz, running on Linux operating system.

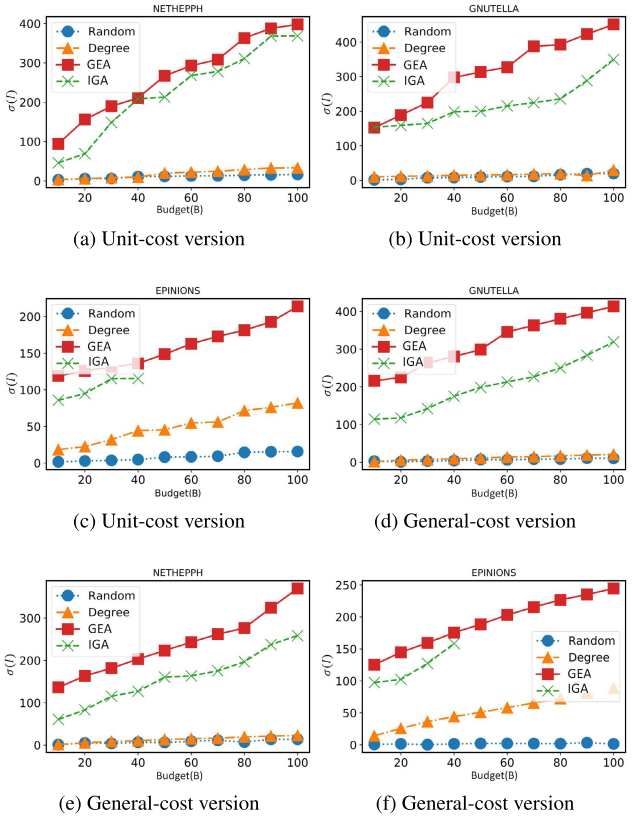
Because it is hard to determine the exact impact weight of node u to v, according to previous researches [5], [16], [36], we set the weight of each edge (u, v) as $w(u, v) = 1/|N_{in}(v)|$. It means that each edge has the same contribution in the activation of a node v, that is $\sum_{u \in N_{in}(v))} w(u, v) = 1$. On the MT-LT model, for each topic, the effect value p_v^i of node v to neighbor nodes and the threshold activation γ_v^i of v according to topic i, $i = 1..q$ are randomly initialized within the range $[0, 1]$. The cost of blocking node $c(v)$, $v \in V$ is randomly initialized within the range $[1.0, 3.0]$. In case costs are identical, we set $c(v) = 1$. The MC simulation method in algorithms is performed to approximately calculate the outcome. The source of misinformation spread S consists of 03 topics ($q = 3$), initially, each topic is randomly contains 100 nodes $|S_1| = 100$, $|S_2| = 100$, $|S_3| = 100$.

The proposed algorithms IGA and GEA are compared with Degree and Random algorithms. For algorithm IGA, 10,000 MC simulations are performed to estimate the outcome of target function $\sigma(\cdot)$. The algorithm GEA (Algorithm 5) is quickly updated with target function value based on depth traversal using tree structure and approximate average denominator on all trees T_i. The algorithm Degree selects all nodes with the highest ranks and adds them to the set of blocked nodes until the total cost for selecting nodes is greater than B, and the algorithm Random selects random nodes within the limited budget B.

B. RESULT

1) EVALUATING ALGORITHMS' EFFICIENCY IN UNIT-COST SETTING

To learn the efficiency of the proposed algorithms, we first conduct some experiments under unit-cost setting. That is, all costs for blocking a node $c(u)$ are 1 for all datasets. The efficiency is measured based on the average outcomes of the diffusion function $\sigma(G, S, A)$ of the formula 4. Fig. 3a, 3b, 3c show the results of all algorithms. When the budget increases, the number of average activated turns increases as well. As we can see, under unit-cost setting, GEA has the best efficiency, followed by IGA and both algorithms outperform Random and Degree with a large margin. In Fig. 3c, we must stop IGA early at budget larger than 40 because this algorithm takes a lot of time.

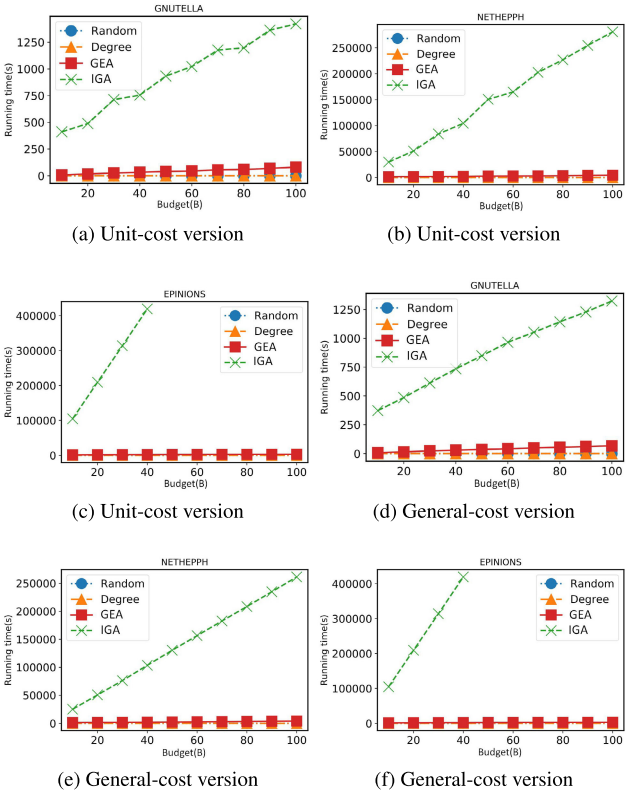


(a) Unit-cost version (b) Unit-cost version

(c) Unit-cost version (d) General-cost version

(e) General-cost version (f) General-cost version

FIGURE 3. (a, b, c) Comparison algorithms in unit-cost version; (d, e, f) Comparison algorithms in general-cost version.



(a) Unit-cost version (b) Unit-cost version

(c) Unit-cost version (d) General-cost version

(e) General-cost version (f) General-cost version

FIGURE 4. (a, b, c) Running time of algorithms in unit-cost version; (d, e, f) Running time of algorithms in general-cost version.

2) EVALUATING ALGORITHMS' EFFICIENCY IN GENERAL-COST SETTING

In this experiment, we compare the algorithms with budget B changing from 0 to 100 and cost of nodes $c(u)$ is evenly distributed within the range [1.0, 3.0]. As can be seen in Fig. 3d, 3e, 3f, both algorithms GEA and IGA outperform Random and Degree algorithms. Algorithm GEA is 1.1 to 2.24 times more efficient than algorithm IGA and up to 121 times more efficient than algorithm Degree in term of the average number of activated turns. The reason is that Degree only uses social network topology attributes but cannot consider the impact process of the source nodes. We stop IGA early at budget larger than 40 on the Epinions network dataset because this algorithm takes a lot of time (longer than 72 hours).

3) COMPARING RUNNING TIME

Finally, we compare the algorithms in running time. Fig. 4a, 4e, 4f and Fig. 4d, 4e, 4f show running time of algorithms on 3 datasets. The running time increases as the budget increases. Random algorithm and Heuristic algorithms are very fast thanks to their simple calculation. Greedy and Random algorithms can run very fast even on big networks. However, GEA algorithm also achieves very competitive running time. The reason is the efficient grouping technique and tree calculation. In all settings, GEA runs faster than IGA up to 196 times. IGA is the slowest algorithm because of the time-consuming in MC simulations.

VI. CONCLUSION

In this paper, we introduce the problem of misinformation blocking with multiple topics spreading on social networks with limited budget. We model the problem as a combination optimization problem based on the LT model with additional requirements of multi-topic and fixed budget for node selection. We propose the MT-LT model to describe the process of multi-topic information spreading by extending the LT model. In this model, information spread is modelled based on different degrees of influence and activation thresholds for each topic. The MMTB problem is formulated by the MT-LT setting. We show that the MMTB problem is NP-hard, the calculation of the objective function is #P-hard and the objective function is monotone and submodular. Based on the monotone and submodular properties of the objective function, we propose an improved greedy algorithm called IGA with approximation ratio $(1 - 1/\sqrt{e})$. Next, we propose an extended algorithm called GEA based on the MT-LT setting by applying a top aggregation method, calculating the sample average and quickly updating the objective function to speed up the algorithm. For those reasons, the proposed algorithm GEA can be applied to medium and large online social networks.

REFERENCES

[1] P. Domm. 2018. *False Rumor of Explosion at White House Causes Stocks to Briefly Plunge; ap Confirms its Twitter Feed Was Hacked.* [Online]. Available: http://www.cnbc.com/id/100646197

[2] H. Allcott and M. Gentzkow. (2019). *Social Media and Fake News in the 2016 Election*. [Online]. Available: https://web.stanford.edu/~gentzkow/research/fakenews.pdf
[3] H. Zhang, M. A. Alim, X. Li, M. T. Thai, and H. T. Nguyen, "Misinformation in online social networks: Detect them all with a limited budget," *ACM Trans. Inf. Syst.*, vol. 34, no. 3, p. 18, 2016.
[4] Y. Zhang and B. A. Prakash, "Scalable vaccine distribution in large graphs given uncertain data," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, Shanghai, China, 2014, p. 1719.
[5] Y. Zhang and B. A. Prakash, "Data-aware vaccine allocation over large networks," *ACM Trans. Knowl. Discovery from Data*, vol. 10, no. 2, pp. 1–32, Oct. 2015.
[6] C. Song, W. Hsu, and M. L. Lee, "Temporal influence blocking: Minimizing the effect of misinformation in social networks," in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, San Francisco, CA, USA, Apr. 2017, p. 541.
[7] (2018). *Swine Flu Frenzy Demonstrates Twitter's Achilles Heel*. [Online]. Available: https//:www.pcworld.com/business center/ article/ 163920/ swine flufrenzy demon strates twitter sachillesheel.html
[8] J. Fan, J. Qiu, Y. Li, Q. Meng, D. Zhang, G. Li, K.-L. Tan, and X. Du, "OCTOPUS: An online topic-aware influence analysis system for social networks," in *Proc. IEEE 34th Int. Conf. Data Eng. (ICDE)*, Paris, France, Apr. 2018, pp. 1569–1572.
[9] Y. Li, D. Zhang, and K.-L. Tan, "Real-time targeted influence maximization for online advertisements," *Proc. VLDB Endowment*, vol. 8, no. 10, pp. 1070–1081, Jun. 2015.
[10] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Washington, DC, USA, 2003, pp. 137–146.
[11] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2001, pp. 57–66.
[12] C. V. Pham, H. M. Dinh, H. D. Nguyen, H. T. Dang, and H. X. Hoang, "Limiting the spread of epidemics within time constraint on online social networks," in *Proc. 8th Int. Symp. Inf. Commun. Technol. (SoICT)*, Nha Trang City, Viet Nam, 2017, pp. 262–269.
[13] M. Kimura, K. Saito, and H. Motoda, "Blocking links to minimize contamination spread in a social network," *ACM Trans. Knowl. Discovery from Data*, vol. 3, no. 2, pp. 1–23, Apr. 2009.
[14] M. Kimura, K. Saito, and H. Motoda, "Solving the contamination minimization problem on networks for the linear threshold model," in *Proc. PRICAI*, Hanoi, Vietnam, Dec. 2008, pp. 977–984.
[15] E. B. Khalil, B. Dilkina, and L. Song, "Scalable diffusion-aware optimization of network topology," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, New York, NY, USA, 2014, pp. 1226–1235.
[16] Y. Zhang, A. Adiga, S. Saha, A. Vullikanti, and B. A. Prakash, "Near-optimal algorithms for controlling propagation at group scale on networks," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3339–3352, Dec. 2016.
[17] B. A. Prakash, H. Tong, N. Valler, M. Faloutsos, and C. Faloutsos, "Virus propagation on time-varying networks: Theory and immunization algorithms," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, Barcelona, Spain, Sep. 2010, pp. 99–114.
[18] H. Tong, B. A. Prakash, C. Tsourakakis, T. Eliassi-Rad, C. Faloutsos, and D. H. Chau, "On the vulnerability of large graphs," in *Proc. IEEE Int. Conf. Data Mining*, Sydney, NSW, Australia, Dec. 2010, pp. 1091–1096.
[19] C. Song, W. Hsu, and M. L. Lee, "Node immunization over infectious period," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, Melbourne, VIC, Australia, 2015, pp. 831–840.
[20] H. Zhang, H. Zhang, X. Li, and M. Thai, "Limiting the spread of misinformation while effectively raising awareness in social networks," in *Proc. CSoNet*, 2015, pp. 35–47.
[21] G. A. Tong, W. Wu, L. Guo, D. Li, C. Liu, B. Liu, and D.-Z. Du, "An efficient randomized algorithm for rumor blocking in online social networks," in *Proc. IEEE Conf. Comput. Commun.*, Atlanta, GA, USA, May 2017, pp. 1–9.
[22] C. Budak, D. Agrawal, and A. El Abbadi, "Limiting the spread of misinformation in social networks," in *Proc. 20th Int. Conf. World Wide Web (WWW)*, Hyderabad, India, 2011, pp. 665–674.
[23] X. He, G. Song, W. Chen, and Q. Jiang, "Influence blocking maximization in social networks under the competitive linear threshold model," in *Proc. SIAM Int. Conf. Data Mining*, Anaheim, CA, USA, Apr. 2012, pp. 463–474.
[24] N. P. Nguyen, G. Yan, and M. T. Thai, "Analysis of misinformation containment in online social networks," *Comput. Netw.*, vol. 57, no. 10, pp. 2133–2146, 2013.
[25] L. G. Valiant, W. Wu, and D. Z. Du, "Distributed rumor blocking with multiple positive cascades," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 2, pp. 468–480, Mar. 2018.
[26] J. Zheng and L. Pan, "Least cost rumor community blocking optimization in social networks," in *Proc. 3rd Int. Conf. Secur. Smart Cities, Ind. Control Syst. Commun. (SSIC)*, Oct. 2018.
[27] M. Farajtabar, J. Yang, X. Ye, H. Xu, R. Trivedi, E. Khalil, S. Li, L. Song, and H. Zha, "Fake news mitigation via point process based intervention," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, Sydney, NSW, Australia, Aug. 2017, pp. 1097–1106.
[28] W. Zhu, W. Yang, S. Xuan, D. Man, W. Wang, and X. Du, "Location-based seeds selection for influence blocking maximization in social networks," *IEEE Access*, vol. 7, pp. 27272–27287, 2019.
[29] C. V. Pham, Q. V. Phu, and H. X. Hoang, "Targeted misinformation blocking on online social networks," in *Proc. ACIIDS*, Dong Hoi City, Vietnam, Mar. 2018, pp. 107–116.
[30] C. V. Pham, Q. V. Phu, H. X. Hoang, J. Pei, and M. T. Thai, "Minimum budget for misinformation blocking in online social networks," *J. Combinat. Optim.*, vol. 38, no. 4, pp. 1101–1127, Nov. 2019.
[31] J. Ma, W. Gao, and K. F. Wong, "Detect rumor and stance jointly by neural multi-task learning," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, New York, NY, USA, Jul. 2016, pp. 3818–3824.
[32] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dallas, TX, USA, Dec. 2013, pp. 1103–1108.
[33] N. Ruchansky, S. Seo, and Y. Liu, "CSI: A hybrid deep model for fake news detection," in *Proc. 2017 ACM Conf. Inf. Knowl. Manage.*, Singapore, Nov. 2017, pp. 797–806.
[34] J. Ma, W. Gao, and K. Wong, "Detect rumor and stance jointly by neural multi-task learning," in *Proc. Companion Web Conf.*, Lyon, France, Apr. 2018, pp. 585–593.
[35] S. Kwon, M. Cha, K. Jung, W. Chen, and Y. Wang, "Prominent features of rumor propagation in online social media," in *Proc. IEEE 13th Int. Conf. Data Mining*, Melbourne, VIC, Australia, Dec. 2013, pp. 40–52.
[36] C. V. Pham, M. T. Thai, H. V. Duong, B. Q. Bui, and H. X. Hoang, "Maximizing misinformation restriction within time and budget constraints," *J. Combinat. Optim.*, vol. 35, no. 4, pp. 1202–1240, May 2018.
[37] L. G. Valiant, "The complexity of enumeration and reliability problems," *SIAM J. Comput.*, vol. 8, no. 3, pp. 410–421, Aug. 1979.
[38] S. Khuller, A. Moss, and J. S. Naor, "The budgeted maximum coverage problem," *Inf. Process. Lett.*, vol. 70, no. 1, pp. 39–45, Apr. 1999.
[39] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Trans. Knowl. Discovery from Data*, vol. 1, no. 1, p. 2, Mar. 2007.
[40] M. Richardson, R. Agrawal, and P. M. Domingos, "Trust management for the semantic Web," in *Proc. ISWC*, Sanibel Island, FL, USA, Oct. 2003, pp. 351–368.
[41] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graphs over time: Densification laws, shrinking diameters and possible explanations," in *Proc. 11th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, Chicago, IL, USA, 2005.

DUNG V. PHAM received the master's degree from Le Quy Don University, in 2012. He is currently pursuing the Ph.D. degree with the Institute of Information Technology, Vietnam Academy of Science and Technology, with a research topic on optimal problems of social networks. He has been working as a Lecturer with the Faculty of Technology and Information Security, People's Security Academy.

GIANG L. NGUYEN received the Ph.D. degree in mathematics from the Vietnam Academy of Science and Technology, in 2012. He is currently an Associate Professor with the Institute of Information Technology, Vietnam Academy of Science and Technology. His research interests include artificial intelligence, data mining, soft computing, and fuzzy computing. He has served as a TPC Chair of many international conferences such as KSE 2017, IJCRS 2019, AICI 2019, MARR 2019, the IEEE-RIVF 2019, and SoICT 2019. He currently serves on the Editorial Board of the *Journal of Computer Science and Cybernetics* (JCC) (Vietnam).

TU N. NGUYEN (Senior Member, IEEE) received the Ph.D. degree in electronics engineering from the National Kaohsiung University of Science and Technology (formerly known as National Kaohsiung University of Applied Sciences), in 2016. He was a Postdoctoral Associate with the Department of Computer Science and Engineering, University of Minnesota Twin Cities, in 2017. Prior to joining the University of Minnesota, he joined the Intelligent Systems Center, Missouri University of Science and Technology, as a Postdoctoral Researcher, in 2016. He is currently an Assistant Professor with the Department of Computer Science, Purdue University Fort Wayne. His research interests include design and analysis of algorithms, network science, cyber-physical systems, and cybersecurity. He has also served as a Technical Program Committee Member of over 70 premium conferences in the areas of networks and communication, such as INFOCOM, GLOBECOM, ICC, and RFID. He has served as a TPC Chair of the ICCASA 2017, NICS 2019, and SoftCOM (25th), a Publicity Chair of iCAST 2017 and BigDataSecurity 2017, and a Track Chair of ACT 2017. He has been serving as an Associate Editor for IEEE Access, since 2019, and the *EURASIP Journal on Wireless Communications and Networking*, since 2017. Since 2017, he has been serving on the Editorial Board of the *Journal of Cybersecurity*, the *Internet Technology Letters*, the *International Journal of Vehicle Information and Communication Systems*, the *International Journal of Intelligent Systems Design and Computing*, and *IET Wireless Sensor Systems*.

CANH V. PHAM received the Ph.D. degree in computer science from the Vietnam National University Hanoi, (VNU). He is currently a Postdoctoral Researcher with the ORLab, Faculty of Computer Science, Phenikaa University. His research interests include information diffusion problems in social networks, combinatorial optimization, and approximation algorithms.

ANH V. NGUYEN received the Ph.D. degree from Kyoto University, in 2012. He has been working as a Senior Researcher with the Institute of Information Technology, Vietnam Academy of Science and Technology. His research interests include machine learning, graph mining, and social network analysis.