IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2018, accepted March 2, 2019, date of publication April 4, 2019, date of current version April 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908412

Knapsack-Based Reverse Influence Maximization for Target Marketing in Social Networks

ASHIS TALUKDER[1], (Member, IEEE), MD. GOLAM RABIUL ALAM[1,2], (Member, IEEE), NGUYEN H. TRAN[3], (Senior Member, IEEE), DUSIT NIYATO[1,4], (Fellow, IEEE), AND CHOONG SEON HONG[1], (Senior Member, IEEE)

[1]Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea
[2]Department of Computer Science and Engineering, BRAC University, Dhaka 1212, Bangladesh
[3]School of Computer Science, The University of Sydney, Sydney, NSW 2006, Australia
[4]School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798

Corresponding author: Choong Seon Hong (cshong@khu.ac.kr)

This work was supported by the MSIT (Ministry of Science and ICT), South Korea, under the Grant Information Technology Research Center Support Program (IITP-2019-2015-0-00742) supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation).

ABSTRACT With the dramatic proliferation in recent years, the social networks have become a ubiquitous medium of marketing and the influence maximization (IM) technique, being such a viral marketing tool, has gained significant research interest in recent years. The IM determines the influential users who maximize the profit defined by the maximum number of nodes that can be activated by a given seed set. However, most of the existing IM studies do not focus on estimating the seeding cost which is identified by the minimum number of nodes that must be activated in order to influence the given seed set. They either assume the seed nodes are initially activated, or some free products or services are offered to activate the seed nodes. However, seed users might also be activated by some other influential users, and thus, the reverse influence maximization (RIM) models have been proposed to find the seeding cost of target marketing. However, the existing RIM models are incapable of resolving the challenging issues and providing better seeding cost simultaneously. Therefore, in this paper, we propose a Knapsack-based solution (KRIM) under linear threshold (LT) model which not only resolves the RIM challenges efficiently, but also yields optimized seeding cost. The experimental results on both the synthesized and real datasets show that our model performs better than existing RIM models concerning estimated seeding cost, running time, and handling RIM-challenges.

INDEX TERMS Influence maximization, reverse influence maximization, target marketing, target marketing cost, social network.

I. INTRODUCTION

Social Network has become the most expected means of communication at present for sharing ideas, views, emotions, news, trends, etc. [1]–[3]. As a result, the number of social network users as well as their network usage are booming day by day and making the social network a very powerful medium for marketing, especially for viral and target marketing. For instance, Facebook crosses the landmark of having two billion active users per month and more than one billion users are active in Wechat per month in 2018 [4]. Again, likewise in real life, people are much influenced by their friends, family members, colleagues, their icon personalities, and even by their favorite brands on social networks [5]. For example, the users could be inspired to travel by any specific airlines or buy brand products promoted by a celebrity on Facebook. This kind of motivation generally spreads among the people in the *word-of-mouth (WoM)* effect in the network [4], [6], [7].

A. BACKGROUND AND MOTIVATION

The Influence Maximization (IM) technique is used to identify the influential users for viral marketing in social networks as well as to estimate the profit earned by the influential seed users. The profit is defined by the maximum number of individuals (nodes) that can be influenced (activated)

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

by the seed users to make some decision (e.g., to grab some products/services) when the seed nodes are activated initially. Several of such IM models are proposed in the literature for various applications including influence maximization [8]–[13], profit maximization [6], [14]–[16], virus, misinformation and rumor management [17], outbreak detection [11], etc. However, almost all of them have a common drawback as they assume the seed nodes are initially activated. Again, in many studies, some incentives, sample products or tickets are offered which might not be enough for seed activation [6], [18]. Since seed nodes influence their followers in the network, the seed nodes might also be activated by some other influential individuals in the same manner. The minimum number of nodes that are required to activate all the seed nodes is termed as seeding cost which is not addressed by most of the state-of-the-art models.

Therefore, Reverse Influence Maximization (RIM) models are introduced to find the seeding cost of viral marketing in which the models diffuse influence reversely as compared to the traditional one [19], [20]. In general, the IM examines who are activated by the seed nodes whereas, the RIM investigates by whom the seed nodes are activated. More specifically, the IM determines the maximum number of nodes that can be activated by a given seed set [21]. On the other hand, the RIM problem aims at finding the least number of nodes that must be activated in order to motivate given seed nodes. The RIM has potential applications in the real-world scenario similar to the IM techniques. For instance, before launching any new product in the market, it is crucial to perform a market analysis and feasibility study called *Cost-Benefit-Analysis (CBA)*. By using the IM in the social network, the vendor may anticipate the tentative profit by estimating the possible product adoption in the social network. In contrary, the RIM can be applied to find the advertisement cost which can be considered to be the seeding cost for target marketing. The authors in [19] identify some RIM challenges which include setting stopping criteria, handling three basic networks components (BNC), NP-Hardness, and insufficient influence. The significant challenges of RIM are to handle the NP-Hardness and to optimize the seeding cost by ensuring approximation rate while keeping the running time faster. However, the existing RIM models are incapable of providing better seeding cost and handling the RIM challenges simultaneously. Thus, we propose a Knapsack-based RIM (KRIM) model which ensures better seeding cost by greedy optimization and resolves the RIM challenges efficiently.

B. OUR CONTRIBUTIONS

In this paper, we consider a scenario of Cost Minimization of Target Marketing in a Social Network, in which seed nodes are targeted for marketing to their huge connections, fans, followers in the social network, and we aim at finding the cost of activating them (making them agree to do such marketing). To do that, we propose a Knapsack-based Reverse Influence Maximization (KRIM) model which estimates the seeding cost of target marketing by disseminating the influence in an order which is opposite to the direction that the influence is diffused in the IM method. We show that the RIM problem is NP-Hard and thus, we employ the Knapsack-based greedy optimization. The KRIM model uses the Linear Threshold (LT) model in reverse order for the node activation process. Again, the node activation process is terminated by the influence decay concept which indicates that the impact of influence is reduced with the hop-distance from the influential node. For instance, an individual has more impact on his/her friends, than his/her friends of friends. The summary of the key contributions is stated below.

1) The proposed KRIM model estimates the optimized seeding cost, as well as addresses the RIM-challenges efficiently.
2) The proposed model employs the greedy approximation technique to manage the NP-Hardness of the problem, and the necessary performance bound (approximation ratio) of the optimization technique is derived.
3) Moreover, we extend the traditional LT model to employ in reverse order and use in the node activation process. At each hop, previously activated nodes are considered first, which enriches the model with more optimized seeding cost.
4) The KRIM model efficiently sets stopping criteria by using the influence decay concept, which is implemented by estimating the cascade influence. Furthermore, the proposed model incorporates the trivial and general case of seed activation whereas, most of the IM models employs only the trivial case.
5) Finally, we evaluate the performance of our model with both the synthesized and real datasets of widely used social networks. The results show that the proposed model outperforms the existing models.

The rest of the paper is organized as follows: a detailed literature review is stated in Section II. We share the RIM problem formulation in Section III and the detailed description of the proposed KRIM model in Section IV. Section V presents the simulation results and performance evaluation of our algorithm while the concluding comments are provided in Section VI.

II. RELATED WORKS

In this section, we study existing methods on Influence Maximization, Profit Maximization, and Reverse Influence Maximization which are relevant to our scenario of cost minimization of target marketing. The detailed analysis of the advantages and disadvantages of the *state-of-the-art* techniques can justify the benefits of our proposed model.

A. INFLUENCE MAXIMIZATION MODELS

Influence Maximization gains huge research interest in recent years in the viral marketing research domain. The IM was first introduced by Domingos and Richardson [10] in 2001 for the online social network and its application was limited to social science and statistics previously.

However, the IM research got a breakthrough in 2003 by Kempe *et al.* [21], with their two classical models such as Linear threshold (LT) and Independent cascade (IC) models. In the LT model, all the nodes are initially considered to be inactive except the seed nodes. Then, all the inactive out-neighbors of the activated nodes are checked whether their integrated incoming influences from activated nodes are no less than some predefined threshold values or not. If so, the node is activated. Then, the out-neighbors of the newly activated nodes are activated in the same fashion. On the other hand, in the IC model, all the nodes are initially inactive as well. However, by assuming that the seed nodes are activated initially, the seed node tries to activate the inactive out-neighbors by a biased coin toss with a specific probability. The process continues until no new node is activated. Both the methods use the greedy approach to maximize the influence and the greedy solution exhibits $(1 - \frac{1}{e}) \approx 63\%$ approximation ratio. Leskovec *et al.* [11] propose a cost-effective lazy forward (CELF) method based on heuristic approximation for outbreak detection. This work is extended by Goyal *et al.* [12] as CELF++ which exhibits 35−55% faster performance than most of the greedy models. The authors, Goyal*et al.* have a series of seminal works in this area, for instance, the learning influence probabilities [22], the simple path technique [13], the databased approach [23], and the IM-based recommender system [15].

A degree discount (DC) heuristic for finding influential nodes is developed by Chen *et al.* [24]. The DC model selects nodes with higher degrees as seeds, and the degree of the chosen node is reduced after node activation which facilitates the greedy method to be more accurate and works faster. Their algorithm improves the accuracy of greedy solution of [21] and the running time of [11] simultaneously. Chen *et al.* [9] also introduce a new paradigm of incorporating negative opinion in influence maximization. Deng *et al.* [25] propose a centrality-based Robust Influence Maximization approach that incorporates various influence functions that are able to handle different uncertainty factors.

For the first time, Rrodriguez and Schölkopf [26] integrate time in the influence maximization process. They formulate a Continuous Time Markov Chain (CTMC)-based approximation algorithm for influence maximization. Most of the state-of-the-art solutions of IM use heuristics and/or time-consuming Monte Carlo (MC) simulation. However, Rrodriguez *et al.* neither use any heuristics nor the MC simulation in CTMC but still, their algorithm gives 20% superior performance to the baseline techniques. Recently, another breakthrough in the IM research is brought by [18] which improves the running time while keeping the approximation rate guaranteed to a better extent of $(1 - \frac{1}{e} - \epsilon)$. It employs an intelligent idea of *stop-and-stare* sampling of dataset yet ensures the quality of the influence estimation at the same time.

However, all of the above studies consider that the seed nodes are activated initially and thus, do not address the seeding cost estimation.

B. PROFIT MAXIMIZATION MODELS

The Influence Maximization techniques are used in many profit maximization applications in social networks. Zhu *et al.* [16] find that influence and profit cannot be maximized together and thus, try to make a balance between them. They introduce a price-aware balanced influence and profit (BIP) model for profit maximization in social networks.

Furthermore, Bhagat *et al.* [6] employ influence maximization to maximize the profit by finding the maximum product adoption in the network. Again, Lu and Lakshmanan [14] extend the work further by incorporating the fact that not only influence but also monetary evaluation affects the adoption of a product. Unlike others, Zhang *et al.* [27] and Du *et al.* [28] perform profit maximization for multiple products whereas, most studies consider only a single product.

In the above studies, the authors trivially offer some free products for seed activation; however, they do not focus on seeding cost estimation in their studies.

C. REVERSE INFLUENCE MAXIMIZATION MODELS

In order to compute the optimized seeding cost, the Reverse Influence Maximization model is introduced by Talukder *et al.* [19]. They propose the Random RIM (R-RIM) and the Randomized Linear Threshold RIM (RLT-RIM) models to solve the RIM problem in which influence is propagated in a backward direction. They also mention some challenges of the RIM problem such as setting stopping condition, handling three basic network components, the hardness of the problem, and insufficient influence. Furthermore, their works are extended in [20] by adding commonality discount. However, none of the above models could resolve the challenging issues properly.

Therefore, in this paper, we propose a Knapsack-based Reverse Influence Maximization (KRIM) model which bestows better seeding cost, as well as handles the challenging issues of RIM efficiently.

III. PROBLEM FORMULATION

For the cost minimization scenario, a social network is given as an input and is represented by a directed graph $G(V, E)$, where the vertex set, V is the set of social network users, and the edge set, E represents the social relations among them. The numbers of nodes and edges are represented by $N = |V|$ and $M = |E|$, respectively. We also denote the out-neighbors and in-neighbors sets of a node v as $n(v)$ and $n^{-1}(v)$, respectively. For each edge $(u, v) \in E$, we calculate the strength of association (influence weight), w_{uv}, which indicates the probability by which the node u influences the node v.

In the Linear Threshold (LT) model, an activation threshold θ_v is also given for each node v. The node v is activated if the combined influence coming from all the active in-neighbors of the node v is no less than the threshold value, θ_v [21], that is,

$$\sum_{u \in n^{-1}(v)} w_{uv} x_u \geq \theta_v, \tag{1}$$

where, x_u indicates whether an in-neighbor u is active or not, *i.e.*,

$$x_u = \begin{cases} 1 & : \text{node } u \text{ is actived,} \\ 0 & : \text{otherwise.} \end{cases} \tag{2}$$

A seed set S of size k is also given as an input. Our goal is to reach these influential users or customers with a view to target marketing of our product or service to a large number of followers in the social network. However, the marketing budget is limited and therefore, it is useful to find a set, denoted by $\Gamma(S)$, of the minimum number of influential users that must be activated in order to subsequently activate influential users. Finally, the seeding cost is defined as $\gamma(S) = |\Gamma(S)|$.

At first, the RIM problem is broken into k subproblems, each for one seed node, as depicted in Fig. 4 and the optimized marginal seeding cost $\gamma(v)$ is computed for each target node, $v \in S$ up to T hops. To do that, the following optimization problem as expressed in (3)-(8) is solved for each hop $t \in T$, and nested with each of activated in-neighbors, u in such a way that the total number of activated nodes, $\gamma(u)$ is minimized.

$$\min \sum_{u \in n^{-1}(v)} x_u \tag{3}$$

$$s.t. \sum_{u \in n^{-1}(v)} w_{uv} x_u \geq \theta_v, \tag{4}$$

$$x_u \in \{0, 1\}, \tag{5}$$

$$w_{uv} \in (0, 1]. \tag{6}$$

The above optimization problem gives the optimal marginal seeding cost set, $\Gamma(v)$ for a seed node v. Similarly, we compute the optimized marginal seeding cost sets $\Gamma(v)$ for all the seed nodes $v \in S$. The optimized marginal cost sets are then, combined to estimate the final optimal seeding cost, $\gamma(S)$ of the whole RIM problem, and is given by,

$$\Gamma(S) = \bigcup_{v \in S} \Gamma(v), \tag{7}$$

$$\gamma(S) = |\Gamma(S)|. \tag{8}$$

*Definition 1 (**RIM Problem**):* Given a social network $G(V, E)$ and a seed set S, the RIM problem is defined by finding the minimum number of nodes, $\gamma(S)$ that must be activated in order to activate all the seed nodes in S.

A. MEETING THE CHALLENGES

Here, we discuss the strategies employed in the proposed KRIM model to resolve the RIM-challenges identified in [19], [20].

1) STOPPING CRITERIA

The KRIM model selects the minimum number of most influential in-neighbors to activate a seed node v at any hop t. Similarly, for each of the chosen in-neighbors, our model chooses the minimum number of most influential in-neighbors at the next hop $t + 1$, and so on, as illustrated in Fig. 4. The existing models estimate the cost up to a fixed number of hops ($T = 2$), which is not proper and adequate.

TABLE 1. Parameter List.

Symbol	Description
$G(V, E)$	Social network
V	Set of social network users
E	Social relationships among users
N	Number of nodes in G, $N = \|V\|$
M	Number of edges in G, $M = \|E\|$
$n(v)$	Set of out-neighbors of v
$n^{-1}(v)$	Set of in-neighbors of v
S	Seed set
k	Size of the seed set S
k_{in}	Number of inactive nodes out of k nodes
$\Gamma(v)$	Marginal seeding cost set of node v
$\Gamma(S)$	Seeding cost set of all the nodes of S
$\gamma(v)$	Marginal seeding cost of v, $\gamma(v) = \|\Gamma(v)\|$
$\gamma(S)$	Seeding cost of the seed set S, $\gamma(S) = \|\Gamma(S)\|$
A_{in}	Active node set that is the input of a hop t
A_{out}	Newly active node set that is the output of a hop t
A_{new}	Nodes that are not previously activated
A_{old}	Nodes that are already activated (previously)
w_{uv}	Social influence of node u to node v
θ_v	Threshold value of the node v
x_u	If u is active, $x_u = 1$ and $x_u = 0$ otherwise
p_c	Cascade influence
T	Total number of hops or iterations
γ	Estimated seeding cost
γ^*	Optimal seeding cost
η	Activation rate
d	Average in-degree in G
C	Complexity of the proposed algorithm

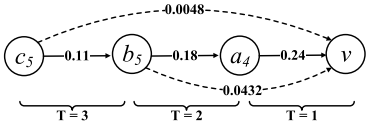


FIGURE 1. Cascade influence (with referred to Fig. 4).

However, we employ the influence decay concept [29], [30] to set the stopping criterion.

*Definition 2 (**Influence Decay**):* The influence of one individual to another individual is deteriorated with the (hop) distance between the individuals. For instance, we have more influence on our friends than on our friends of friends. □

In the proposed KRIM model, the influence decay concept is implemented by estimating the cascade influence of the multi-hop-distanced nodes.

*Definition 3 (**Cascade Influence**):* The influence weight of a node u to another node v is cascade influence if $(u, v) \notin E$. In Fig. 1, the dotted lines indicate the cascade influence; whereas, the solid lines indicate the direct influence or just influence. □

If p_t is the influence weight at any hop $t \in T$, the cascade influence, p_c of total T hops is calculated by the influence

decay function given in (9).

$$p_c = \prod_{t=1}^{T} p_t \tag{9}$$

The concept in (9) is partially adopted from the live edge concept from [13] (pp. 3) and we name it as cascade influence. In order to adopt the concept in the KRIM model, let us assume that p_t is the minimum influence weight of the activated in-neighbors u to activate the node v at every hop t, and then, p_t is computed as,

$$p_t = \{\min w_{zu} | z \in A_{out}, u \in A_{in} \text{ at any hop } t\}, \tag{10}$$

where, A_{in} denotes the list of nodes that are the input to any hop t, and A_{out} indicates the list of nodes that are activated in the current hop t, as shown in Fig. 5.

The logic behind taking the minimum influence weight among the influence weights of the activated in-neighbors at every hop is that the cost computation spans to the minimum possible hops, and eventually, the cost remains minimum.

When the value of the cascade influence, p_c becomes negligible (e.g., $p_c < 10^{-6}$) at some hop t, it indicates that the individuals at the hop t do not have any significant influence on the seed node v, and the KRIM model terminates at the hop t. We consider that the nodes after t-hop have the insignificant influence upon v, and can be ignored.

*Example 1 (**Influence Decay and Cascade Influence**):* In Fig. 1 and Fig. 4, at the first hop, $A_{in} = \{v\}$ and their activated in-neighbors, $A_{out} = \{a_2, a_4\}$. Then, $p_1 = \min(w_{a_2v}, w_{a_4v}) = \min(0.37, 0.24) = 0.24$, by (10). Similarly, we have,

$$\begin{aligned} &\text{At} t = 2, \quad A_{in} = a_2, a_4, \; A_{out} = b_2, b_4, \; p_2 = 0.18 \\ &\text{At} t = 3, \quad A_{in} = b_2, b_4, \; A_{out} = c_4, c_5, \; p_3 = 0.11 \\ &\qquad\qquad \cdots \quad \cdots \end{aligned} \tag{11}$$

Therefore, the cascade influence of the node c_5 to the node v at 3-hop distance is estimated by (9) as $p_c = 0.11 \times 0.18 \times 0.24 = 0.0048$. Finally, when the value of p_c becomes negligible (e.g., $p_c < 10^{-6}$) at any hop t, the algorithm terminates at the hop t. □

2) DIFFERENT CASES

Three Basic Network Components (BNC) are considered in existing models [19] such as a seed node with no in-neighbors (Case A), one hop in-neighbors (Case B), and multiple in-neighbors (Case C). However, Case C is the combination of Case A and Case B. Thus, Case C is a redundant BNC, and hence, the KRIM considers only the first two cases, which are shown in Fig. 2.

a: HANDLING THE TRIVIAL CASE (CASE A)

Case A, depicted in Fig. 2 (a), is a trivial case. When this case happens in the network, we can offer free samples suggested by [6] and thus, the cost is estimated as, $\gamma(v) = |\Gamma(v)| = |\{v\}| = 1$. Most of the Influence Maximization and Profit Maximization models incorporate the trivial case only by offering sample product to the seed users and thus, ignore the general case. Our KRIM model is well designed that it includes both the trivial and general cases. When the values of $\forall x_u = 0$, and $x_v = 1$ for all $u, v \in \Gamma(v)$, then, the trivial case occurs and the KRIM model sets, $\gamma(v) = 1$.

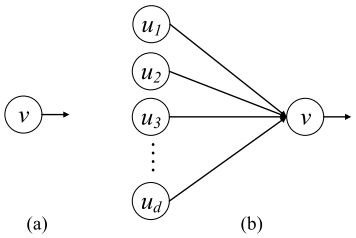


FIGURE 2. Basic Network Components (BNC): (a) Trivial Case and (b) General Case.

b: HANDLING THE GENERAL CASE (CASE B)

In viral or target marketing, the seed nodes influence and activate their friends and followers (*i.e.*, out-neighbors) in a cascade manner. However, most of the existing algorithms do not include the general case which indicates that the seed nodes could also be influenced and activated by their friends and followees (*i.e.*, in-neighbors) in a reverse cascade fashion. The proposed KRIM model incorporates the general case as well as the trivial case. The general case occurs when $\forall x_u \in \{0, 1\}$, and $x_v = 1$ for all $u, v \in \Gamma(v)$. The inclusion of this fact is one of the vital contributions of this paper.

Case B, in which the seed node v has only one level of in-neighbors, is the basic unit of computation as shown in Fig. 2 (b). The KRIM uses Knapsack-based LT model to find the optimized cost for this case.

The whole social network can be viewed as a combination of the multiple instances of these two cases, and thus, the KRIM considers only two cases out of three cases considered in the existing models.

3) INSUFFICIENT INFLUENCE

There may exist an unfavorable situation in the network that all the in-neighbors, $u \in n^{-1}(v)$ may not have enough combined influence to activate a node v. The event is termed as the insufficient influence, which is mathematically expressed as,

$$\sum_{u \in n^{-1}(v)} w_{uv} x_u < \theta_v, \quad \exists v \in V. \tag{12}$$

The first two parameters depend on the inherent network structure (how many friends/followers a person might have), which is intractable. Thus, there remains no alternative other than setting the threshold values to some smaller value [31]. If the insufficient influence arises even with a lower threshold, our algorithm returns all the in-neighbors as seeding cost, *i.e.*, $\gamma(v) = |\Gamma(v)| = |n^{-1}(v) \cup \{v\}|$, and assumes that the node is activated.

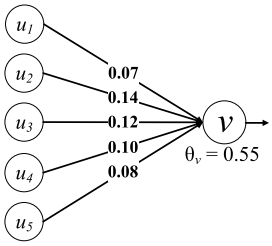


FIGURE 3. The insufficient influence. The aggregated influence ($\sum w_{uv} = 0.51$) of all the in-neighbors is less than the threshold value ($\theta_v = 0.55$). Therefore, all the in-neighbors together can not activate the node v.

*Example 2 (**Insufficient Influence**):* In Fig. 3, the node v has insufficient influence since $\sum_{u \in n^{-1}(v)} w_{uv} = 0.07 + 0.14 + 0.12 + 0.10 + 0.08 = 0.51 < \theta_v = 0.55$. Here, the aggregated influence of all the in-neighbors of v is less than the threshold of v, and thus, they cannot activate v all together. We denote this situation as a *true positive (TP) insufficient influence* or just insufficient influence.

In case of the node v in Fig. 3, the KRIM model sets all the in-neighbors in the seeding cost set, $\Gamma(v)$ and in-degree as the seeding cost, *i.e.*, $\gamma(v) = |\Gamma(v)| = |\{u_1, u_2, u_3, u_4, u_5\}| = 5$. It is also assumed that the node v is activated with the influence of all its in-neighbors. □

4) NP-HARDNESS OF THE PROBLEM

Here, we discuss the hardness of the RIM problem under the proposed KRIM model.

Theorem 1: The RIM problem under KRIM model is NP-Hard.

Proof: Let us consider the Knapsack problem [20],

$$\max \sum_{u=1}^{n} x_u p_u \tag{13}$$

$$s.t. \sum_{u=1}^{n} x_u w_u \leq m, \tag{14}$$

$$x_u \in \{0, 1\}, \tag{15}$$

where n is the number of items, m is the Knapsack size, w_u is the weight of the item u, p_u is the profit of item u, and x_u is defined in (2).

In order to transform the Knapsack problem into the RIM problem, let us consider the Knapsack size m is the counterpart of the threshold value θ_v in the RIM problem. The item weights w_u in the Knapsack problem can be considered to be equivalent to the influence weights w_{uv} in the RIM problem. Again, let us consider the profit $\sum p_u x_u$ of the selected items in the Knapsack problem as the counterpart of the seeding cost $\sum x_u$ of the selected nodes (according to RIM formulation, each node contributes unit cost) in the RIM problem. Finally, let us replace the objective function of the RIM problem stated in (3) as:

$$\max \frac{1}{\sum_{u \in n^{-1}(v)} x_u}. \tag{16}$$

TABLE 2. Reducing Knapsack into RIM.

No.	Parameters in Knapsack	Parameters RIM
1	Knapsack size, m	Threshold, θ_v
2	Item Weights, w_u	Influence weight, w_{uv}
3	Profit, $\sum p_u x_u$	Seeding cost, $\sum x_u$ (according to the RIM formulation, every activated node incurs unit cost)
4	**Objective:** maximize the profit taking item with higher unit price (greedy) until the Knapsack size is reached.	**Objective:** minimize the seeding cost taking highest influence weight (greedy) until the threshold is reached.
5	**Termination:** When the aggregated weight of the selected items reaches the Knapsack size.	**Termination:** When the aggregated influence weight of the selected in-neighbor nodes reaches the node threshold.

Now, the profit maximization in the Knapsack problem is technically the same as the cost minimization in the RIM problem. The summary of their parameters comparison is given in Table 2.

Thus, the Knapsack problem, which is a well-known NP-Hard problem [32], is reduced to the RIM problem and hence, the RIM problem under KRIM model is also NP-Hard. □ ■

Therefore, we offer a Knapsack-based greedy approximation model (KRIM) to solve the RIM problem.

B. THE KRIM MODEL

The proposed KRIM model first, estimates the optimized marginal seeding cost $\gamma(v)$, of all the seed nodes $v \in S$ and then, aggregates them to find the seeding cost $\gamma(S)$ of the whole problem. The optimized marginal seeding cost, in return, is computed in two steps: a diffusion model is used for the node activation process, and an optimization technique is used to minimize the cost.

1) THE LT DIFFUSION MODEL AND GREEDY KNAPSACK OPTIMIZATION

The Linear Threshold (LT) model which is generally applied in a forward manner in IM problems to determine the nodes that can be activated by a seed node v when it is activated initially. However, we modify the LT model to employ in a retrograde manner to determine which nodes are required to activate the seed node v.

To estimate the optimized marginal seeding cost set $\Gamma(v)$, we iterate up to T hops of in-neighbors of v. The determination of the value of T is discussed earlier in Section III (A). Let A_{in} denotes the list of nodes that are the input to any hop t, contains the activated nodes in the previous hop and A_{out} represents the list of nodes that are activated in the current hop t, as shown in Fig. 5.

We start the process of finding the optimized marginal cost set, $\Gamma(v)$ with the following initializing at the first

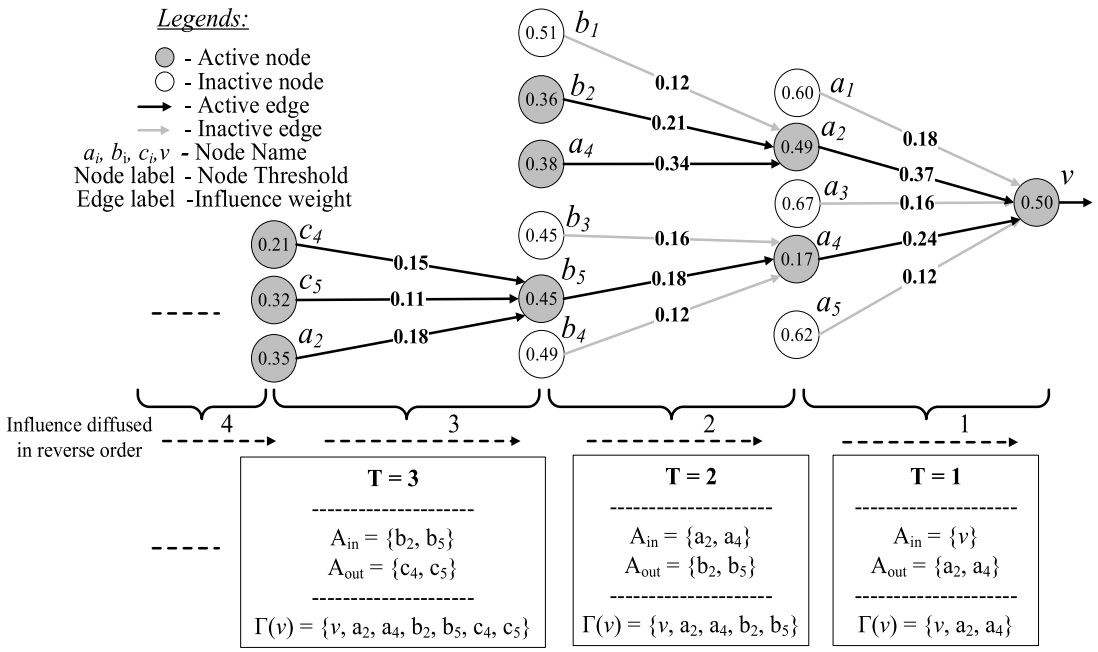


FIGURE 4. The working principles of the KRIM model.

hop ($T = 1$),

$$A_{in} = \{v\}, \tag{17}$$

$$A_{out} = \emptyset. \tag{18}$$

Then, the nodes that are required to be activated for each node $u \in A_{in}$, are estimated by applying the LT model in a reverse order and selecting the most influential in-neighbors greedily so that the cost remains minimum.

2) TRIVIAL CASE

If the trivial case (Case A) happens, the KRIM model returns the node u itself, *i.e.*,

$$A_{out} = \{u\}. \tag{19}$$

3) GENERAL CASE

Furthermore, if the general case (Case B) is encountered for a node $u \in A_{in}$, the previously activated in-neighbor nodes in $n^{-1}(u)$ are considered first to activate u. Thereafter, if necessary, the inactive in-neighbors in $n^{-1}(u)$ are considered next to activate each node $u \in A_{in}$.

At any hop $t \in T$, if $\Gamma(S)$ is the set of activated nodes by previously processed seed nodes, and $\Gamma(v)$ contains the nodes that are activated by the current seed node v. Then, the sets of already activated nodes and inactive nodes are given, respectively by,

$$A_{old} = n^{-1}(u) \cap (\Gamma(S) \cup \Gamma(v)), \tag{20}$$

$$A_{new} = n^{-1}(u) \setminus (\Gamma(S) \cup \Gamma(v)). \tag{21}$$

Firstly, the node z with the highest influence weight w_{zu} is selected in a greedy (Knapsack) manner from the set of already activated nodes A_{old}, in order to activate u with threshold θ_u. Secondly, if necessary (if u is not activated by the already activated in-neighbors), the new inactive in-neighbors are chosen to activate u greedily, as expressed by (22) and (23), respectively.

$$A_1 = \sum_{z \in A_{old}} \left[\arg\max_{u \in A_{in}} w_{zu} \right] \geq \theta_u, \tag{22}$$

$$A_2 = \sum_{z \in A_{new}} \left[\arg\max_{u \in A_{in}} w_{zu} \right] \geq \theta_u. \tag{23}$$

Every time that the node z is selected, the influence weight w_{zu} is aggregated and is checked by (1), whether the node u is activated or not. Then, the output A_{out} of the hop t is updated with the newly activated nodes as,

$$A_{out} = (A_1 \cup A_2) \setminus (\Gamma(S) \cup \Gamma(v)). \tag{24}$$

Here, the already activated nodes (both by previous and current seed nodes) are also excluded since they are already explored in the cost estimation. Again, the optimized marginal seeding cost set is populated as,

$$\Gamma(v) = \Gamma(v) \cup A_{out}. \tag{25}$$

Finally, the output A_{out} of the current hop t is forwarded as the input A_{in} to the next hop $t + 1$ and the whole process is continued up to T hops. Similarly, after estimating the optimized marginal seeding cost set, $\Gamma(v)$ for all $v \in S$, the final optimal seeding cost is computed by (7) and (8). The process is illustrated in Fig. 1 to Fig. 5, and in Example 3.

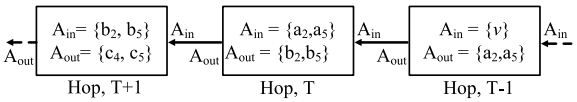


FIGURE 5. Hop-based node activation process in the KRIM model.

C. INFLUENCE WEIGHT ESTIMATION

Many techniques are available in the literature to estimate the influence weight w_{uv}, for instance, degree centrality, PageRank [21], credit distribution [22], assigning a constant value [33], and Trivalency model [34]. However, we adopt the most widely used *degree centrality* technique which computes w_{uv} as,

$$w_{uv} = \frac{1}{|n^{-1}(v)|}, \tag{26}$$

where, $|n^{-1}(v)|$ is the in-degree of node v provided that the normalization property holds, *i.e.*,

$$\sum_{u \in n^{-1}(v)} w_{uv} \leq 1. \tag{27}$$

*Example 3 (**Working Principles of KRIM Model**):* The primary goal of the KRIM model is to determine which nodes jointly activate the seed node v, and the calculation of $\gamma(v)$ for a single seed node v is explained here, as depicted in Fig. 4.

At the first hop, $t = 1$, we assume that the currently activated node set, $A_{in} = \Gamma(v) = \{v\}$. At the beginning, there is no previously activated node in $n^{-1}(v)$ and thus, we select newly activated nodes, $A_{out} = \{a_2, a_4\}$ with maximum influence weights by (23) to activate the seed node v. This gives, $\Gamma(v) = \{v, a_2, a_4\}$.

At $t = 2$, we have, $A_{in} = \{a_2, a_4\}$. To activate a_2, already activated node a_4 from $n^{-1}(a_2) = \{a_4, b_1, b_2\}$ is considered first. Then b_2 is selected greedily from new inactive nodes $\{b_1, b_2\}$. On the other hand, $\{b_5\}$ is selected to activate a_4. Here, $\{a_4, b_2, b_5\}$ are required to activate nodes in A_{in}. However, a_4 is explored in previous hop. Thus, we have, $A_{out} = \{b_2, b_5\}$, and $\Gamma(v) = \{v, a_2, a_4, b_2, b_5\}$.

Similarly, at $t = 3$, we have $A_{in} = \{b_2, b_5\}$. Note that the node b_5 suffers from insufficient influence, and we include all its in-neighbors $\{a_2, c_4, c_5\}$ as the cost. On the other hand, node b_2 is the trivial case as explained in Fig. 2(a) and thus, only b_2 is included as cost. However, a_2 and b_2 are explored previously and thus, $A_{out} = \{c_4, c_5\}$. Finally, we have, $\Gamma(v) = \{v, a_2, a_4, b_2, b_5, c_4, c_5\}$.

Note that for any seed node v_m, we also consider the already activated nodes by previous seed nodes $\{v_1, v_2, \ldots, v_{m-1}\}$ and the nodes activated by current seed node v_m to achieve the least possible seeding cost.

Again, in this example, some repeated nodes are given in some hops since this is a graph, not a tree. Nonetheless, we also show the repeated nodes in a tree structure just for better realization. □

Algorithm 1 The SEEDING Module

Input: $G(V, E)$, v, $\Gamma(S)$
Result: *active*, $\Gamma(v)$

1 $\Gamma(v) := \{v\}$;
2 *active* $:= 0$; /* **Inactive** */
3 $A_{in} := \{v\}$;
4 $p_c := 1.0$, $t := 1$;
5 **while** *true* **do**
6 **if** $p_c < 10^{-6}$ **then**
7 *break*; /* **Stopping criteria** */
8 **end**
9 $A_{out} := \emptyset$;
10 $p_a := p_n := 1.0$;
11 **for** $u \in A_{in}$ **do**
12 $A_{old} := n^{-1}(u) \cap (\Gamma(S) \cup \Gamma(v))$; /* **Already active node** */
13 **if** $A_{old} \neq \emptyset$ **then**
14 A_1, *active*, $p_a :=$ ACTIVATE(A_{old}, u); /* **From already activated nodes** */
15 **end**
16 $A_{new} := n^{-1}(u) \setminus (\Gamma(S) \cup \Gamma(v))$; /* **Inactive new nodes** */
17 **if** *active* $== 2$ & $A_{new} \neq \emptyset$ **then**
18 A_2, *active*, $p_n :=$ ACTIVATE(A_{new}, u); /* **From New nodes** */
19 **end**
20 **if** *active* $== 0$ **then**
21 *active* $:= 3$; /* **Reporting insufficient influence** */
22 **end**
23 **end**
24 $A_{out} := (A_1 \cup A_2) \setminus (\Gamma(S) \cup \Gamma(v))$;
25 $\Gamma(v) := \Gamma(v) \cup A_{out}$; /* **Optimized marginal seeding cost set** */
26 $t := t + 1$; /* **Next hop,** $t + 1$ */
27 $p_t := \min(p_a, p_n)$;
28 $p_c := p_c * p_t$; /* **Cascade influence** */
29 $A_{in} := A_{out}$; /* **For next hop** */
30 **end**
31 *return* $\Gamma(v)$, *active*;

D. THE KRIM ALGORITHM

The KRIM algorithm, composed of three modules such as KRIM module (Algorithm 3), the SEEDING module (Algorithm 1), and the ACTIVATE module (Algorithm 2), estimates the optimized seeding cost for a given seed set S. The calling sequence of different modules is illustrated in the control-flow diagram shown in Fig. 6.

The KRIM module estimates the optimized marginal seeding cost set $\Gamma(v)$, for all the seed nodes $v \in S$, in lines $3 - 16$. It handles the trivial case (Case A) in lines $4 - 7$ and the general case (Case B) in lines $8 - 14$. Again, in the general case, the KRIM module calls the SEEDING module for

Algorithm 2 The ACTIVATE Module

Input: A, u
Result: $active, A_{out}, p_{min}$
1 $A_{out} := \emptyset$, $influence := 0.0$, $active := 0$;
2 **while** $A \neq \emptyset$ **do**
3 | **if** $inf\_sum \geq \theta_v$ **then**
4 | | $active := 2, break$; /* **Active** */
5 | **end**
6 | Select $z \in A$ with max w_{zu};
7 | $influence := influence + w_{zu}$;
8 | $A_{out} := A_{out} \cup \{z\}$;
9 | $A := A - \{z\}$;
10 **end**
11 $p_t := \{\min w_{zu} | z \in A_{out}, u \in A_{in}\}$;
12 *return* $A_{out}, active, p_t$;

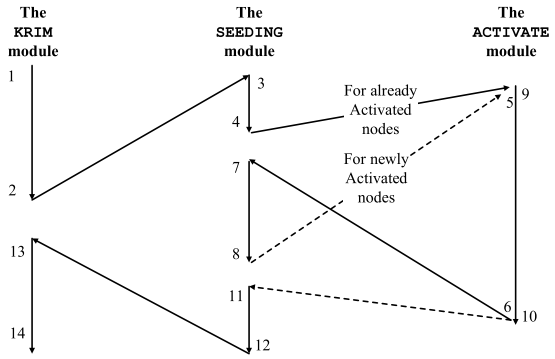


FIGURE 6. Control-flow diagram of different modules of the proposed model.

each v in line 10 to estimate the optimized seeding cost set, $\Gamma(v)$. Then, it combines all the seeding cost set of all the seed nodes to estimate the final seeding cost set $\Gamma(S)$, in line 15. Finally, it finds the seeding cost $\gamma(v)$, in line 18. The KRIM module also keeps track of the number of active and inactive (due to the insufficient influence) seed nodes in lines 12 and 17, respectively.

The SEEDING module estimates $\Gamma(v)$ up to T hops iteratively. The terminating condition is set in lines $6-8$. This module first tries to activate any node by its already activated in-neighbors in lines $12-15$. Secondly, it includes the new active in-neighbors (if necessary) in lines $16-19$. In both cases, it calls the ACTIVATE module, which greedily selects the most influential in-neighbors (with the highest w_{uv} to reduce the cost). It reports the insufficient influence in lines $20-22$ and estimates cascade influence in line 28. Finally, the optimized marginal seeding cost set $\Gamma(v)$, is computed in line 25.

E. THE PERFORMANCE BOUND

Generally, a greedy model provides a feasible solution with significantly improved running time. However, under some specific conditions, it can provide the best solution. Similarly,

Algorithm 3 The KRIM Module

Input: $G(V, E), S$
Result: $\gamma(S), \Gamma(S)$
1 $\Gamma(S) := \emptyset$;
2 $active\_count := 0$; /* **Active seed nodes** */
3 **for** *each* $v \in \mathcal{S}$ **do**
4 | **if** $n^{-1}(v) == \emptyset$ **then**
5 | | $\Gamma(v) := \{v\}$, $active := 1$;
6 | | *return* $\Gamma(v)$; /* **Trivial case** */
7 | **end**
8 | **else**
9 | | $\Gamma(v) := \emptyset$;
10 | | $\Gamma(v)$, $active := \text{SEEDING}(G, v, \Gamma(S))$; /* **General case** */
11 | | **if** $active == 2$ **then**
12 | | | $active\_count := active\_count + 1$; /* **No. of active seed nodes** */
13 | | **end**
14 | **end**
15 | $\Gamma(S) := \Gamma(S) \cup \Gamma(v)$;
16 **end**
17 $k_{in} := k - active\_count$; /* **No. of inactive (insufficient influence) seed nodes** */
18 $\gamma(S) := |\Gamma(S)|$; /* **Final seeding cost** */
19 *return* $\gamma(S), k_{in}$;

the KRIM method provides an optimum solution with some particular constraints as well.

1) FEASIBLE SOLUTION AND APPROXIMATION RATIO

A special case of the greedy feasible solution and the approximation ratio are stated in the following two theorems.

Theorem 2: The KRIM model ensures optimum result when there is no overlapping in the diffusion process, *i.e.*, if

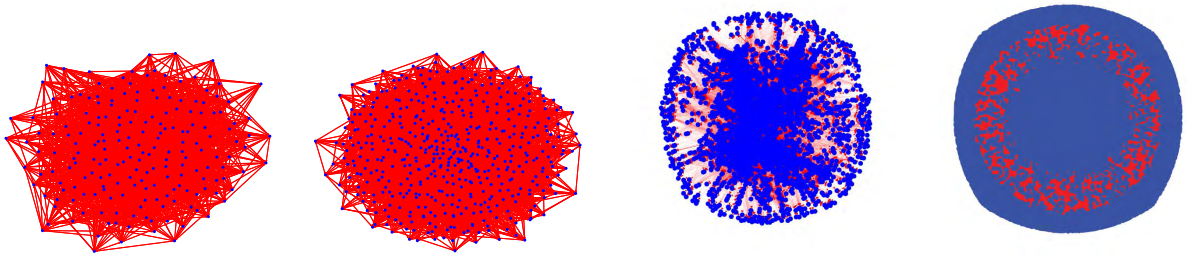
$$\Gamma(v_1) \cap \Gamma(v_2) = \emptyset, \quad \forall v_1, v_2 \in S, \tag{28}$$

and there is no insufficient influence in the network, *i.e.*, if

$$\sum_{u \in n^{-1}(v)} w_{uv} x_u \geq \theta_v, \quad xx \forall v \in S. \tag{29}$$

Proof: If every node v has the property that its in-neighbors have enough aggregated influence to activate it, then, it is evident that the KRIM model must activate all the given seed nodes.

Again, if there is no overlapping in any stage of the node activation process, *i.e.*, if $\Gamma(v_1) \cap \Gamma(v_2) = \emptyset, \forall v_1, v_2 \in S$, there would be no already-activated node. Therefore, the most influential nodes will be always chosen in line 6 of the ACTIVATE module and hence, the cost must be the minimum. □



(a) Erdös Rényi (b) Barabási Albert (c) Facebook (d) Twitter

FIGURE 7. Networks used for simulation and performance evaluation. a) The Erdös Rényi network having $N = 200$ and $M = 2073$, b) The Barabási Albert network having $N = 500$ and $M = 4900$, c) Facebook network having $N = 4,093$ and $M = 88234$ and d) Twitter network having $N = 81306$ and $M = 1768,149$ (Legends: blue dots = nodes and red lines = edges).

Again, when the algorithm cannot ensure the optimum solution due to the use of the greedy approach, it is necessary to analyze the approximation ratio.

Theorem 3: The KRIM algorithm is a 2 - *approximation* algorithm, that is,

$$\gamma \leq 2\gamma^* \tag{30}$$

Proof: In Theorem 1, we prove that the Knapsack problem can be reduced to the RIM problem under KRIM algorithm. Let r be the index in the sorted influence weight list and the nodes up to r are selected to activate a node v. Then, we claim that influence weights are taken,

$$\underbrace{w_{u_1v}}_{1} + \underbrace{w_{u_2v}}_{2} + \cdots + \underbrace{w_{u_{r-1}v}}_{\text{r-1}} + \underbrace{w_{u_rv}}_{\text{r}} (\geq \theta_v), \tag{31}$$

and the incurred cost is bounded by the optimal cost [35], and is given by,

$$x_1 + x_2 + \cdots + x_r \geq \gamma^*, \tag{32}$$

where γ^* is the optimal value of cost, and any solution is a feasible solution with $\gamma^* \leq \gamma$. We set the cost of nodes, $x_1 = x_2 = \cdots = x_r = 1$ and $x_i = 0$ for all $r < i \leq d$ which is a feasible solution, and cannot be improved since we always select the most influential in-neighbor u with maximum w_{uv}. Therefore,

$$x_1 + x_2 + \cdots + x_r = \gamma \geq \gamma^* \tag{33}$$

Again, according to Hochbaum [36], the left hand side in (33) must be at best $2\gamma^*$. Thus, we get,

$$\gamma \leq 2\gamma^* \qquad \square$$

2) COMPLEXITY

The complexity of the KRIM model is expressed by the number of nodes that it processes. If d is the average in-degree in G, the expected running time of our algorithm is given by:

$$C \leq \underbrace{k}_{\text{seeds}} \underbrace{T}_{\text{hops}} (\underbrace{d^2}_{\text{already}} + \underbrace{d^2}_{\text{new}}) \approx O(kTd^2). \tag{34}$$

TABLE 3. Datasets Description.

Dataset Type	Network Name	Nodes	Edges
Synthesized	Erdös Rényi	200	2,071
	Barabási Albert	500	4,900
Real	Facebook	4,039	88,234
	Twitter	81,306	1,768,149

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the KRIM model for two synthesized datasets and two real datasets of popular social networks. We analyze the estimated seeding cost as well as the required running time by comparing the KRIM model with those of the existing models.

A. DATASET

For the simulation, we use real datasets of popular social networks as well as synthesized datasets generated randomly. We prepare two synthesized datasets by generating networks randomly. The first one is the *Erdös Rényi* network generated randomly with parameters, $(200, 0.1)$[1], having 200 nodes and 5,094 edges as depicted in Fig. 7 (a). The second one is the *Barabási Albert* graph generated randomly with parameters, $(500, 10, 100)$[2], having 500 nodes and 4,900 edges as shown in Fig. 7 (b). We collect datasets of two most popular social networks named *Facebook*[3], and *Twitter*[4], from the Stanford large network dataset collection [37]. The Facebook is a *friendship* network, whereas, the Twitter is *follower-followee* network. For instance, a link (u, v) between two users indicates that the users u and v are friends on Facebook. On the other hand, a link (u, v) in the Twitter network indicates that a Twitter user, u follows the user v on the Twitter. The Facebook and the Twitter networks are depicted in Fig. 7 (c) and Fig. 7 (d), respectively. The summary of all the datasets is presented in Table 3.

[1]Number of nodes = 200, and link probability = 0.1.
[2]Number of nodes = 500, initial links = 10, and Seed = 100.
[3]https://snap.stanford.edu/data/egonets-Facebook.html
[4]https://snap.stanford.edu/data/ego-Twitter.html

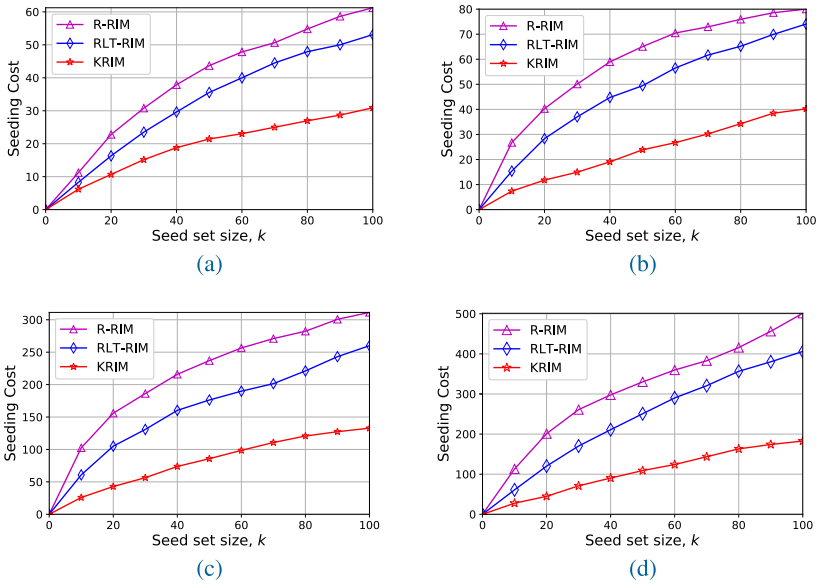


FIGURE 8. Seeding cost for $k = 1$ to 100 (with fixed $\theta_v = 0.10$), for a) Erdös Rényi, b) Barabási Albert, c) Facebook, and d) Twitter dataset.

B. EXPERIMENT SETUP

We perform the simulation of our algorithm on an Intel(R) Core(TM) i3-4150 CPU @ 3.50GHz 3.50GHz machine with 8GB RAM, using Python codes.

The influence weights, w_{uv} are computed by using the *degree centrality* [21] technique which is commonly used in the social network analysis. The values of k are varied from 1 to 100 in the experiment while keeping the threshold value fixed at, $\theta_v = 0.10$ [21]. This small value is taken as the threshold value, θ_v to avoid insufficient influence [20], [31]. Again, the threshold values vary as, $\theta_v = 0.0$ to 0.50, in the experiment with fixed seed size $k = 50$. Generally, the seed set S is supposed to be given. However, we generate randomly for our simulation. Since the problem is NP-hard, the Monte Carlo [38] simulation is applied. The algorithms are executed $10,000$ times on all the datasets, and the average of each parameter is taken for the comparative study. We compare the proposed KRIM algorithm with the following existing RIM models [19]:

1) R-RIM: The main algorithm is designed to compute the seeding cost for only the first two hops; however, we redesign the model for T hops. At each hop, to activate a node, we randomly select a number of in-neighbors which are then used as input for the next hop iteration.
2) RLT-RIM: This model also estimates the seeding cost up to the first two hops and thus, we extend it for T hops as well. At each hop, to activate a node, we randomly choose an in-neighbors and compare the exaggerated influence with the threshold of the node. If the node is activated, the process halts. Otherwise, the process continues until the node is activated. The same process continues for the chosen in-neighbors in the next hop iteration.

C. PERFORMANCE ANALYSIS

Here, we discuss the simulation results of the proposed KRIM model with a comparative analysis with the existing models concerning the estimated seeding cost, the running time, and the efficient handling of the RIM-challenges. The salient features of the result analysis are that all the empirical results are explained and supported with theoretical and probabilistic analysis along with appropriate figures and examples.

1) SEEDING COST

Fig. 8 and Fig. 9 depict the seeding cost estimated by different models for both the synthesized datasets (Erdös Rényi and Barabási Albert) and real datasets (Facebook and Twitter).

In Fig. 8, the cost is calculated for different seed set size, $1 \leq k \leq 100$, and with fixed threshold value, $\theta_v = 0.10$. Compared with the R-RIM and RLT-RIM models, the KRIM algorithm returns the optimized seeding cost, which is on an average $1.5 - 2$ times lower for the Erdös Rényi dataset, $1.5-2$ times lower for the Barabási Albert dataset, $2-3$ times lower for the Facebook dataset, and $1.5 - 2.5$ times lower for the Twitter dataset. Therefore, the proposed KRIM model has remarkably better performance than those of the existing models for different values of k, for all the four datasets due to the use of the (Knapsack) greedy in-neighbor selection.

Furthermore, we analyze the effect of different threshold values on the estimated seeding cost in Fig. 9 for all the datasets. We employ threshold values, $0 \leq \theta_v \leq 0.5$ and the fixed size seed set with $k = 50$, for the both real and the synthesized datasets. The simulation results show that the estimated seeding cost of the proposed KRIM model is significantly lower than those of the existing R-RIM and RLT-RIM models for all the datasets. For instance, as compared to the existing models, the estimated cost of the KRIM algorithm is

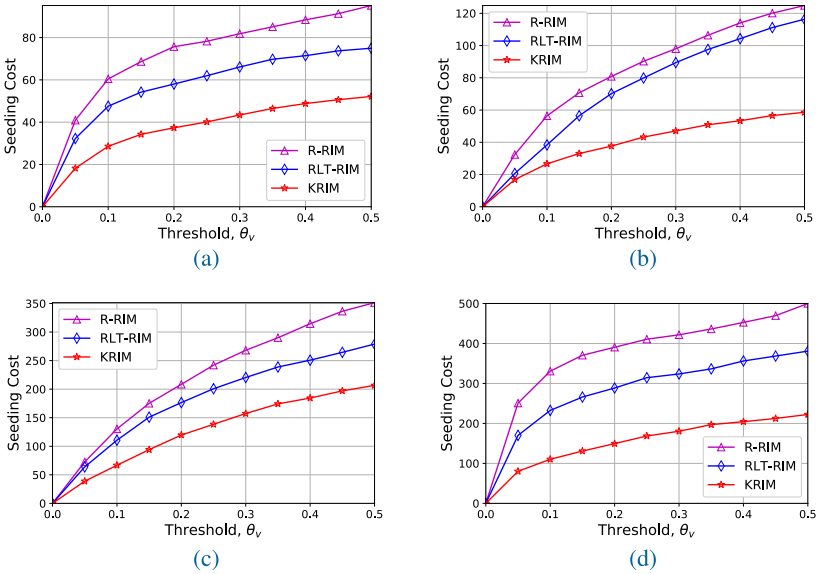


FIGURE 9. Seeding cost for $\theta_v = 0$ to 0.40 (with fixed $k = 50$), for a) Erdös Rényi, b) Barabási Albert, c) Facebook, and d) Twitter dataset.

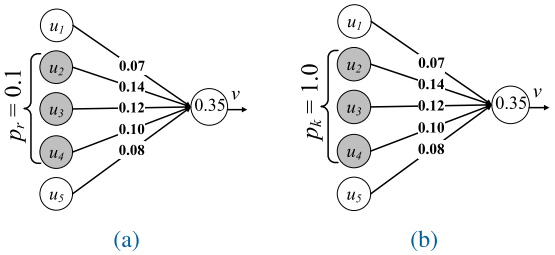


FIGURE 10. The Probabilistic Analysis of cost optimization. (a) Random node selection by the R-RIM and RLT-RIM models (b) Greedy node selection by the KRIM model.

around $30-45\%$, $40-50\%$, $30-40\%$, and $45-55\%$ lower for the Erdös Rényi, Barabási Albert, Facebook, and Twitter datasets, respectively. Thus, the KRIM model also exhibits superior performance for variable threshold values and all the datasets.

This significant performance of the KRIM method is the contribution of the greedy Knapsack method used in the in-neighbors selection and the node activation process. Therefore, in both the cases of different k-values and the θ_v-values, the proposed KRIM model evidently outperforms the existing RIM models for all four datasets.

a: THEORETICAL ANALYSIS OF COST OPTIMIZATION

In the following Example 4, we provide a theoretical analysis of the cost optimization achieved by the proposed KRIM model with random models such as R-RIM and RLT-RIM [19].

*Example 4 (**Theoretical Analysis of Seeding Cost**):* We assume that r out of d in-neighbors are required to activate v, depending upon the threshold θ_v of the node v. In Fig. 10, for both the cases of random and greedy in-neighbor selection, $d = 5$ and $r = 3$, *i.e.*, at least three nodes must be taken to activate the node v. With the given influence weight setup, the most influential set $\{u_2, u_3, u_4\}$ is the optimal solution. The selection of any node except these three nodes will increase the estimated cost. In case of the R-RIM and RLT-RIM models in Fig. 10 (a), the in-neighbor, u_i is selected randomly, whereas, in Fig. 10 (b), the in-neighbor, u_i is selected greedily in the KRIM model. Here, according to the working principles of the RLT-RIM technique, the probability of selecting the in-neighbors, $\{u_2, u_3, u_4\}$ is,

$$p_r(\{u_2, u_3, u_4\}) = \frac{1}{{}^dC_r} = \frac{1}{{}^5C_3} = \frac{1}{10} = 0.10. \quad (35)$$

However, according to the working principles of the KRIM model, the probability of selecting the in-neighbors, $\{u_2, u_3, u_4\}$ is,

$$p_k(\{u_2, u_3, u_4\}) = 1. \quad (36)$$

Since the KRIM model chooses nodes greedily, it selects the node with the highest influence each time. Therefore, we can conclude that the KRIM model can provide (local) optimal result in 90% of all cases as compared to the R-RIM and RLT-RIM methods. Moreover, in the random algorithms, if any node other than $\{u_2, u_3, u_4\}$ is selected, the cost is elevated. This can happen, since every node has equal probability (in Fig. 10 (a), $\frac{1}{d} = \frac{1}{5}$) to be selected as cost. □

2) RUNNING TIME

The running time of the proposed model and the existing models is shown in Fig. 11 and Fig. 12, for variable seed set size ($0 \leq k \leq 100$) and different threshold values ($0.0 \leq \theta_v \leq 0.5$), respectively, for both the real and the synthesized datasets.

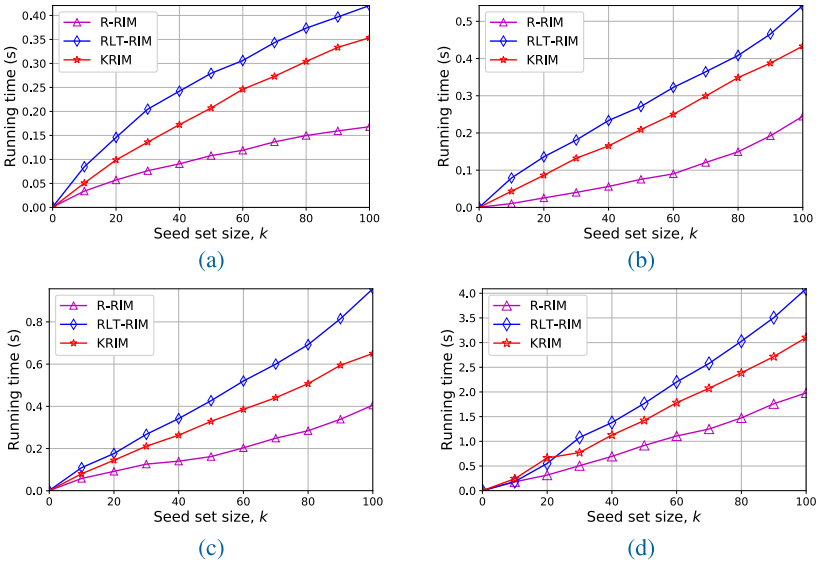


FIGURE 11. Running time for $k = 1$ to 100 (with fixed $\theta_V = 0.10$), for a) Erdös Rényi, b) Barabási Albert, c) Facebook, and d) Twitter dataset.

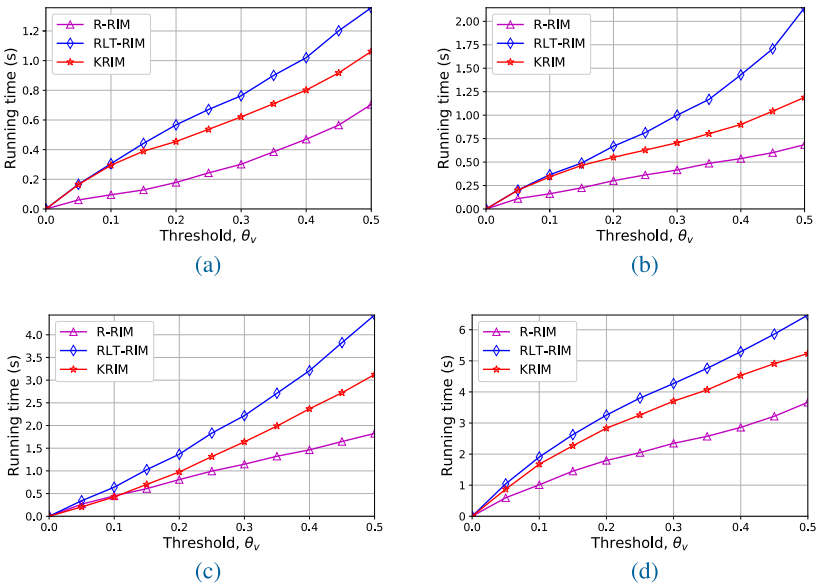


FIGURE 12. Running time for $\theta_V = 0$ to 0.50 (with fixed $k = 50$), for a) Erdös Rényi, b) Barabási Albert, c) Facebook, and d) Twitter dataset.

For running time estimation, the threshold value is fixed at $\theta_v = 0.1$. However, the seed size, k is varied from 0 to 100 in Fig. 11. On the other hand, in Fig. 12, the seed size is fixed at $k = 50$, and the threshold value θ_v, is varied from 0 to 0.50.

In Fig. 11, the proposed algorithm exhibits reasonable running time for all the datasets. The figures unveil that the running time of our model is sandwiched between that of two existing models. As compared to the RLT-RIM model, the KRIM model is around 15% and 20% faster for all the four datasets. However, the proposed model is slightly slower than the R-RIM model. The R-RIM model requires less time for its execution since it activates the nodes randomly while calculating the seeding cost. However, the R-RIM returns bad quality seeding cost, which is 15−50% higher than that of the proposed model, depending upon the various datasets used in the simulation.

Next, with the variable threshold values, $0.0 \leq \theta_v \leq 0.50$, the proposed KRIM model exhibits the same pattern of running time which lies in between the running time of the R-RIM and RLT-RIM models as shown in Fig. 12. Thus, the KRIM model yields an optimized seeding cost with reasonable running time.

*Example 5 (**Theoretical Analysis of Running Time**):* Again, consider Fig. 10 (a), the R-RIM model takes a random number of in-neighbors, and it does neither aggregates nor compares with the threshold value. Thus, the

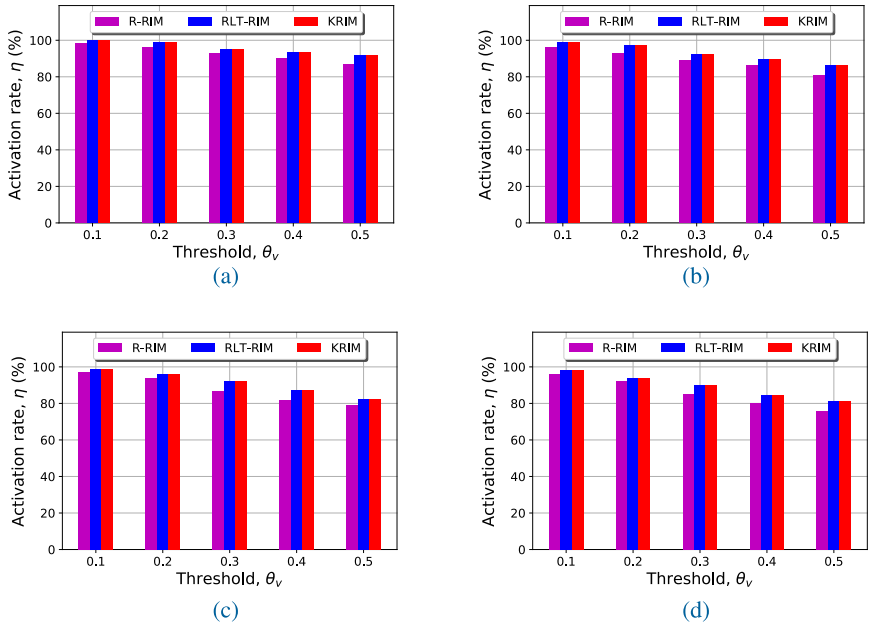


FIGURE 13. Activation rate for $\theta_v = 0$ to 0.50 (with fixed $k = 100$), for a) Erdös Rényi, b) Barabási Albert, c) Facebook, and d) Twitter dataset.

R-RIM technique requires the least amount of running time. On the other hand, in both the RLT-RIM and KRIM algorithms, every time a node u is chosen, its influence weight w_{uv} is aggregated and checked by (1) for the activation of v. As a result, these two models require a higher running time than that of the R-RIM model.

However, the RLT-RIM model chooses in-neighbor nodes randomly. On the other hand, the KRIM model selects the most influential nodes each time. Therefore, the RLT-RIM model has a higher probability of selecting more in-neighbor nodes to reach the threshold value of a target node. Therefore, the RLT-RIM method consumes more time as well as returns the higher seeding cost as compared to the KRIM model. □

D. TIME-COST TRADE-OFF

The R-RIM exhibits the fastest running time among all the three models. However, the R-RIM model suffers from two significant drawbacks. Firstly, the model is not cost-effective since it randomly selects the in-neighbor nodes to activate a node, and thus, has the highest cost margin. Secondly, the R-RIM algorithm has the worst *activation rate* among all the discussed models as illustrated in Fig. 13. The activation rate is defined as the percentage of the total activated seed nodes $(k - k_{in})$ out of total seed nodes k as stated in (37). For instance, in the Fig. 10 (a), if the R-RIM model selects the in-neighbors $\{u_1, u_4, u_5\}$, the seed node v might remain inactive even though the node v does not have insufficient influence $(\sum w_{uv} = 0.07 + 0.14 + 0.12 + 0.10 + 0.08 = 0.51 > \theta_v = 0.35)$. We define this situation as *false positive (FP) insufficient influence*. Therefore, the R-RIM model has very low efficiency in spite of better running time.

Furthermore, the RLT-RIM is neither the cost-effective nor the time-efficient compared with the proposed model. It has a longer running time even though it cannot ensure the optimal seeding cost compared with the KRIM model. Finally, the KRIM model requires intermediary running time. However, it is the most cost-effective RIM model.

1) ANALYSIS OF RIM CHALLENGES

Here, we discuss how effectively the KRIM model resolves the challenges of the RIM problem.

a: STOPPING CONDITION

The proposed KRIM model resolves the challenge of setting up a terminating criterion more efficiently as compared to the existing R-RIM and RLT-RIM models by employing the influence decay concept given in (9). The influence decay concept [29], [30] and cascade influence concept [13] illustrated in Fig. 1. The influence decay concept is easy to implement by (9) and (10), more realistic, and more logical to set up a stopping condition than to a fixed number of hops, which is used in the R-RIM and RLT-RIM models.

b: DIFFERENT CASES

The proposed model considers only two cases, *i.e.*, the trivial case and the general case as shown in Fig. 2. On the other hand, the existing R-RIM and RLT-RIM models employ three cases such as Case A, Case B, and Case C. As a result, the existing models become more complex. Thus, the proposed model is more straightforward than the existing models.

c: INSUFFICIENT INFLUENCE

If any seed node suffers from the insufficient influence effect in the cost estimation process, we assume that the seed node v is activated with the influence of all its in-neighbors.

However, in reality, the seed node v remains inactivated. Therefore, we measure the efficiency of any RIM model with the number of active and inactive seed nodes. If k_{in} nodes out of k seed nodes remain inactive after the node activation process of any RIM model, then the *activation rate* η, is given by,

$$\eta = \frac{k - k_{in}}{k} \times 100\%. \tag{37}$$

A node may remain inactive due to either the insufficient influence (true positive) of the limitation of the RIM model (false positive insufficient influence) as we discussed earlier.

Fig. 13 depicts the estimated *activation rates* η, of all the discussed RIM models for all the datasets. Here, the rate is estimated with different threshold values, $0.0 \leq \theta_v \leq 0.5$ and the seed size fixed at $k = 100$. The experimental results show that the insufficient influence effect is elevated with the increased values of threshold that is the expected result.

Again, the empirical result also shows that the RLT-RIM and the KRIM models have the same activation rate for all the datasets. However, the activation rate of the R-RIM model is lower than that of both the RLT-RIM and KRIM models.

*Example 6 (**Theoretical Analysis of Activation Rate**):* The major difference between the working principles of the proposed KRIM model and the RLT-RIM model is that the KRIM model selects the higher influential in-neighbor nodes one by one greedily. On the other hand, the RLT-RIM model chooses randomly. Due to this difference, the KRIM offers lower seeding cost than that of the RLT-RIM model. Note that although the KRIM model returns more optimized seeding cost than that of the RLT-RIM model, both the models exhibit the same performance in the case of insufficient influence. For instance, in Fig. 3, for the node v, both the models selects the same seeding cost set, $\Gamma(v) = n^{-1}(v) \cup \{v\} = \{v, u_1, u_2, u_3, u_4, u_5\}$, except the node selection sequence. For example, the KRIM algorithm selects the sequence as $\{v, u_2, u_3, u_4, u_5, u_1\}$ with the decreasing influence weight. On the other hand, the RLT-RIM selection sequence is random, e.g., $\{v, u_4, u_1, u_3, u_5, u_2\}$. After selecting the same cost set (probably in a different order), both the algorithms report the real occurrence of insufficient influence, which is also known as the true positive (TP) insufficient influence case. Therefore, the activation rate η is the same for RLT-RIM and KRIM models.

On the other hand, the R-RIM model not only faces true positive insufficient influence but also results in false positive insufficient influence in many cases. In other words, when the true insufficient influence (TP) happens, the R-RIM model must report it. However, the R-RIM model might report insufficient influence even when there is no insufficient influence for a node v in the network. For instance, in Fig. 10, the node v has no insufficient influence (since $\sum w_{uv} \geq \theta_v$). However, the R-RIM model may randomly select the in-neighbors $\{u_1, u_3\}$ and report that node v has insufficient influence which is a false positive case. Thus,

$$\text{for R-RIM, } k_{in} \text{ has } = \text{TP effect} + \text{FP effect}, \tag{38}$$

$$\text{for RLT-RIM, } k_{in} \text{has } = \text{TP effect}, \tag{39}$$

$$\text{and, for KRIM, } k_{in} \text{has } = \text{TP effect}. \tag{40}$$

Therefore, the associated activation rates of the three models have the relation $\eta_R \leq \eta_{RLT} = \eta_K$, which validates the empirical result. □

d: HARDNESS OF THE PROBLEM

The KRIM model uses greedy Knapsack technique to solve the NP-Hardness of the problem, and therefore, returns better-optimized seeding cost than those of the existing models. The existing R-RIM and RLT-RIM models are incapable of handling this issue effectively since they do not employ any optimization method such as the greedy Knapsack method. In the proposed model, we also derive the approximation ratio of the optimization technique that makes the KRIM model more promising.

V. CONCLUSIONS

In this paper, we have proposed a Knapsack-based Reverse Influence Maximization (KRIM) model to estimate the seeding cost for target marketing in social networks. The proposed model employs the Linear Threshold (LT) model in reverse order in the node activation process. Moreover, we have used the greedy Knapsack technique to optimize the seeding cost. Therefore, we can provide a theoretical performance bound of the proposed model which makes our model more promising. Our model resolves the RIM-challenges more efficiently and provides the most economical seeding cost simultaneously. Furthermore, we have integrated a practical feature of viral marketing in our model such as influence decay concept which indicates that the influence subsides with the (hop) distance from the influential node. Finally, we have evaluated the performance of our proposed model using two synthesized datasets and two real datasets, and the simulation results show that our model outperforms the existing models.

REFERENCES

[1] R. Mohamadi-Baghmolaei, N. Mozafari, and A. Hamzeh, "Trust based latency aware influence maximization in social networks," *Eng. Appl. Artif. Intell.*, vol. 41, pp. 195–206, May 2015.
[2] Y. Hu, R. J. Song, and M. Chen, "Modeling for information diffusion in online social networks via hydrodynamics," *IEEE Access*, vol. 5, pp. 128–135, 2017.
[3] L. Alsuwaidan and M. Ykhlef, "Information diffusion predictive model using radiation transfer," *IEEE Access*, vol. 5, pp. 25946–25957, 2017.
[4] F. Ye, J. Liu, C. Chen, G. Ling, Z. Zheng, and Y. Zhou, "Identifying influential individuals on large-scale social networks: A community based approach," *IEEE Access*, vol. 6, pp. 47240–47257, 2018.
[5] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *ACM Trans. Web*, vol. 1, no. 1, p. 5, 2007.
[6] S. Bhagat, A. Goyal, and L. V. Lakshmanan, "Maximizing product adoption in social networks," in *Proc. 5th ACM Int. Conf. Web Search Data Mining*, 2012, pp. 603–612.
[7] J. J. Brown and P. H. Reingen, "Social ties and word-of-mouth referral behavior," *J. Consum. Res.*, vol. 14, no. 3, pp. 350–362, 1987.
[8] N. Barbieri, F. Bonchi, and G. Manco, "Topic-aware social influence propagation models," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 81–90.
[9] W. Chen *et al.*, "Influence maximization in social networks when negative opinions may emerge and propagate," in *Proc. SDM*, vol. 11, 2011, pp. 379–390.

[10] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 57–66.
[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.
[12] A. Goyal, W. Lu, and L. V. Lakshmanan, "CELF++: Optimizing the greedy algorithm for influence maximization in social networks," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 47–48.
[13] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "SIMPATH: An efficient algorithm for influence maximization under the linear threshold model," in *Proc. IEEE 11th Int. Conf. Data Mining (ICDM)*, Dec. 2011, pp. 211–220.
[14] W. Lu and L. V. S. Lakshmanan, "Profit maximization over social networks," in *Proc. IEEE 12th Int. Conf. Data Mining (ICDM)*, Dec. 2012, pp. 479–488.
[15] A. Goyal and L. V. Lakshmanan, "RecMax: Exploiting recommender systems for fun and profit," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 1294–1302.
[16] Y. Zhu, Z. Lu, Y. Bi, W. Wu, Y. Jiang, and D. Li, "Influence and profit: Two sides of the coin," in *Proc. IEEE 13th Int. Conf. Data Mining*, Dec. 2013, pp. 1301–1306.
[17] A. Talukder *et al.*, "Rumors in the social network: Finding the offenders using influence maximization," in *Proc. Korean Comput. Congr. (KCC)*, 2015, pp. 1214–1216.
[18] H. T. Nguyen, M. T. Thai, and T. N. Dinh, "Stop-and-stare: Optimal sampling algorithms for viral marketing in billion-scale networks," in *Proc. Int. Conf. Manage. Data*, 2016, pp. 695–710.
[19] A. Talukder *et al.*, "An approach of cost optimized influence maximization in social networks," in *Proc. 19th Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2017, pp. 354–357.
[20] A. Talukder, M. G. R. Alam, N. H. Tran, and C. S. Hong, "A cost optimized reverse influence maximization in social networks," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Apr. 2018, pp. 1–9.
[21] 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*, 2003, pp. 137–146.
[22] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 241–250.
[23] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *Proc. VLDB Endowment*, vol. 5, no. 1, pp. 73–84, Sep. 2011.
[24] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 199–208.
[25] X. Deng, Y. Dou, T. Lv, and Q. V. H. Nguyen, "A novel centrality cascading based edge parameter evaluation method for robust influence maximization," *IEEE Access*, vol. 5, pp. 22119–22131, 2017.
[26] M. G. Rodriguez and B. Schölkopf. (2012). "Influence maximization in continuous time diffusion networks." [Online]. Available: https://arxiv.org/abs/1205.1682
[27] H. Zhang, H. Zhang, A. Kuhnle, and M. T. Thai, "Profit maximization for multiple products in online social networks," in *Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Apr. 2016, pp. 1–9.
[28] N. Du, Y. Liang, M. F. Balcan, and L. Song. (2013). "Budgeted influence maximization for multiple products." [Online]. Available: https://arxiv.org/abs/1312.2164
[29] S. Feng, X. Chen, G. Cong, Y. Zeng, Y. M. Chee, and Y. Xiang, "Influence maximization with novelty decay in social networks," in *Proc. AAAI*, 2014, pp. 37–43.
[30] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck. (2014). "Distance-based influence in networks: Computation and maximization." [Online]. Available: https://arxiv.org/abs/1410.6976
[31] A. Talukder, M. G. R. Alam, A. K. Bairagi, S. F. Abedin, H. T. Nguyen, and C. S. Hong, "Threshold estimation models for influence maximization in social network," in *Proc. Korean Inst. Inf. Sci. Eng. (KIISE)*, 2016, pp. 888–890.
[32] E. Horowitz and S. Sahni, *Fundamentals of Computer Algorithms*. Rockville, MD, USA: Computer Science Press, 1978.
[33] Q. Yu, H. Li, Y. Liao, and S. Cui, "Fast budgeted influence maximization over multi-action event logs," *IEEE Access*, vol. 6, pp. 14367–14378, 2018.
[34] A. Talukder, A. K. Bairagi, D. H. Kim, and C. S. Hong, "Reverse path activation-based reverse influence maximization in social networks," *J. Korean Inst. Inf. Sci. Eng.*, vol. 45, no. 11, pp. 1203–1209, 2018.
[35] A. Krause and D. Golovin, "Submodular function maximization," in *Tractability: Practical Approaches to Hard Problems*, vol. 3, no. 19. Cambridge, U.K.: Cambridge Univ. Press, 2012, pp. 1–8.
[36] D. S. Hochbaum, *Approximation Algorithms for NP-Hard Problems*. Boston, MA, USA: PWS Publishing, 1996.
[37] J. Leskovec and A. Krevl. (Jun. 2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. http://snap.stanford.edu/data
[38] B. Liu, G. Cong, D. Xu, and Y. Zeng, "Time constrained influence maximization in social networks," in *Proc. IEEE 12th Int. Conf. Data Mining*, Dec. 2012, pp. 439–448.

ASHIS TALUKDER (S'17–M'18) received the B.S. and M.S. degrees in computer science and engineering from the University of Dhaka, Bangladesh, where he has been an Assistant Professor with the Department of Management Information Systems (MIS), since 2009. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University, South Korea. His research interests include social networks, influence maximization, network optimization, and data mining. He is a member of the IEEE Communication Society (IEEE ComSoc), Association for Information Systems (AIS), Bangladesh Chapter, and Internet Society, Bangladesh Chapter. He is also a member of the Korean Institute of Information Scientists and Engineers (KIISE).

MD. GOLAM RABIUL ALAM (S'15–M'17) received the B.S. degree in computer science and engineering, the M.S. degree in information technology, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2017, where he was a Postdoctoral Researcher with the Computer Science and Engineering Department, from 2017 to 2018. He is currently an Associate Professor with the Computer Science and Engineering Department, BRAC University, Dhaka, Bangladesh. His research interests include healthcare informatics, mobile cloud and Edge computing, ambient intelligence, and persuasive technology. He is a member of the IEEE IES, CES, CS, SPS, CIS, and ComSoc. He is also a member of KIISE and received several best paper awards from prestigious conferences.

NGUYEN H. TRAN (S'10–M'11–SM'18) received the B.S. degree in electrical and computer engineering from the Ho Chi Minh City University of Technology, in 2005, and the Ph.D. degree in electrical and computer engineering from Kyung Hee University, in 2011, where he was an Assistant Professor with the Department of Computer Science and Engineering, from 2012 to 2017. Since 2018, he has been with the School of Information Technologies, The University of Sydney, where he is currently a Senior Lecturer. His research interests include applying analytic techniques of optimization, game theory, and machine learning to cutting-edge applications, such as cloud and mobile-edge computing, data centers, resource allocation for 5G networks, and the Internet of Things. He received the Best KHU Thesis Award in engineering, in 2011, and several best paper awards, including IEEE ICC 2016, APNOMS 2016, and IEEE ICCS 2016. He has been the Editor of the IEEE TRANSACTIONS ON GREEN COMMUNICATIONS AND NETWORKING, since 2016.

DUSIT NIYATO (M'09–SM'15–F'17) received the B.Eng. degree from the King Mongkuts Institute of Technology Ladkrabang (KMITL), in 1999, and the Ph.D. degree in electrical and computer engineering from the University of Manitoba, Canada, in 2008. He is currently a Full Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include green communication, the Internet of Things (IoT), and sensor networks.

CHOONG SEON HONG received the B.S. and M.S. degrees in electronic engineering from Kyung Hee University, Seoul, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree from Keio University, in 1997. In 1988, he joined KT, where he worked on Broadband Networks as a member of the Technical Staff. Since 1993, he has been with Keio University, Japan. Since 1999, he has been a Professor with the Department of Computer Science and Engineering, Kyung Hee University. His research interests include the future Internet, ambient intelligent technology, wireless networks, network security, and network management. He is also a member of the ACM, IEICE, IPSJ, KIISE, KICS, and KIPS. He has served as a Program Committee Member and an Organizing Committee Member for International conferences, such as NOMS, IM, APNOMS, and ICOIN.