An Approach of Cost Optimized Influence Maximization in Social Networks

Ashis Talukder, Md. Golam Rabiul Alam, Anupam Kumar Bairagi, Sarder Fakhrul Abedin, Md Abu Layek, Hoang T. Nguyen, and Choong Seon Hong
Department of Computer Engineering, Kyung Hee University, South Korea
Email: {ashis, robi, anupam, saab0015, layek, nguyenth, cshong}@khu.ac.kr

***Abstract*—Social networks have gained huge research interest, especially in viral marketing due to their rapid boom in the past years. It is very crucial to identify the influential users in the social networks for viral and target marketing. Influence maximization (IM) problem estimates such influential users in the social networks. With an initial seed set, the IM finds a maximum number of nodes that can be activated in the network under some diffusion models $e.g.$ Linear Threshold model or Independent Cascade model. But previous works in this field have not studied about the minimum cost, termed as opportunity cost (OC), to motivate those seed nodes. In this work, we define a novel Reverse Influence Maximization (RIM) problem to determine the opportunity cost of influence maximization. Employing the influence propagation in opposite order, the RIM determines the minimum number of nodes that must be activated in order to motivate a set of target nodes. We propose Random RIM (R-RIM) and Randomized Linear Threshold RIM (RLT-RIM) models to tackle the RIM problem. We also perform a simulation to evaluate the performance of the algorithms using two real world datasets. The result shows that the proposed models determine the optimized opportunity cost with faster running time margin.**

***Index Terms*—influence maximization, reverse influence maximization, RIM, viral marketing, opportunity cost, social network.**

I. Introduction

With the rapid proliferation of the number of sites and their usage, social networks have become the ultimate way of connecting people and sharing information, news, trends etc. Hence, social networks have become the attractive medium of marketing and field of research as well [1]. In the last one and a half decade, Influence Maximization (IM) has gained a huge interest in the social network research. The IM problem finds a small seed set of users such the total number of activated nodes is maximized in the network. It is also called viral marketing where influence spreads in *word-of-mouth* effect [2]. It mirrors the human behavior of real life scenario that people always consult with the family members, friends, colleagues, or other experts before taking any decision $viz.$ any purchasing decision [3].

The IM problem has various attractive applications $e.g.$ profit maximization or maximizing product adaptation [4], [5], rumor spread and detection [6], contaminations and outbreak detection [7] etc. Thus, most researches have been conducted in these directions but estimating the minimum cost of influence maximization has not been addressed deeply. Many researchers have just offered free sample products or free tickets of a concert to the seed users [4], [5] in their research. The approach is not rational since those influential users are human and they also might be motivated by some other people. But none of the studies has addressed the cost of activating those influential users. Therefore, we compute the minimum opportunity cost to activated the influential target users.

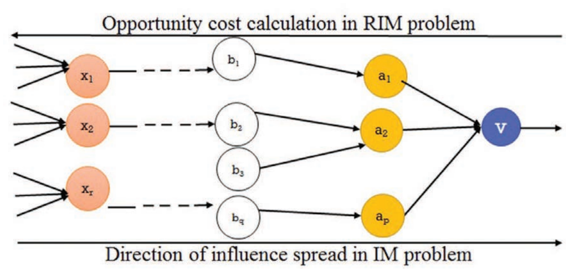


Figure 1: Basic working principle of RIM compared to IM

In this research, we have introduced a novel Reverse Influence Maximization (RIM) problem to estimate the opportunity cost [8] which is defined by the minimum number of nodes that must be activated in order to activate the given set of target nodes in a social network. The IM problem finds a set of seed users that maximizes the spread of influence in the network [2]. On the other hand, RIM determines the minimum number of in-neighbors that are required to activated a given set of target nodes in the network. Thus, the RIM works in the opposite manner as compared to IM problem as stated in the Fig. 1. This is the rationale of the naming of RIM. We propose two models namely, Random RIM (R-RIM) and Randomized Linear Threshold RIM (RLT-RIM) models to solve RIM problem. The RLT-RIM is based on the classical Linear threshold model [2] where the R-RIM is a purely random model. We have evaluated the performance of these models using two real datasets [9].

This research was supported by the Ministry of Science and ICT (MSIT), Korea, under the Grand Information Technology Research Center support program (IITP-2017-2015-0-00742), supervised by the Institute for Information & communications Technology Promotion (IITP)".
*Dr. CS Hong is the corresponding author.

II. Existing Study

The nominal work has been carried out by Kempe et al. [2] who proposed two classical models named *Linear Threshold (LT)* and *Independent Cascade (IC)* models in 2003.

Leskovec et al. [7] have proposed the Cost Effective lazy Forward (CELF) model which outperforms 700 times faster than the standard greedy algorithms and 90% optimal as well. Chen et al. [10] have devised Maximum Influence in Arborescences with Negative opinion (MIA-N) model which exhibits faster running time with the same approximation ratio as the greedy model. Goyal et al. [11] have formulated a path based algorithm that shows the better result than many existing models. A heuristic degree discount method has been introduced in [12] and it has improved the accuracy of classical models [2] and the running time of CELF model simultaneously.

The authors in [12] have formulated a linear time and LT based Local directed acyclic graph (LDAG) model which is scalable to extremely large networks. The authors in [4] have considered multiple products rather than a single product in profit maximization. Generally, previous works have involved only single product but the fact is that any company does not produce only one product but multiple products. Nguyen et al. [13] have integrated multiple social networks in their work with a fact that multiple networks can support each other to propagate influence among the networks.

But none of the above researches have addressed the problem of determining the opportunity cost. In this paper, we propose two such models to estimate the opportunity cost of influence maximization in social networks along with two solution models.

III. Problem formulation

Consider a social network represented by a graph $G(V, E)$, where each vertex is a user and each edge is a social relationship between two users. The influence weight w_{uv} is defined as the probability of user u to influence the user v. According to the LT model [2], a node v is activated if $\sum_{u\in n^{-1}(v)} w_{uv}x_u \geq \theta_v$, where $n^{-1}(v)$ is the in-neighbor set of V and x_u indicates whether a node u is activated ($x_u = 1$) as an in-neighbor of any target node or not ($x_u = 0$). For a given target set S of k influential users, the RIM aims at finding the opportunity cost $\sigma(S)$.

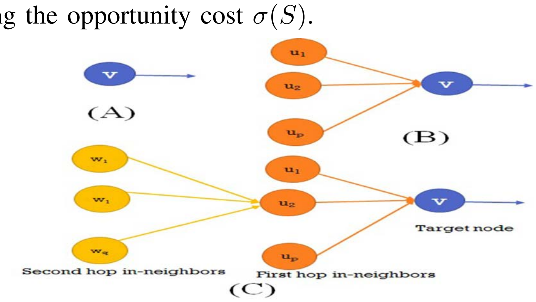


Figure 2: Basic network components of RIM

The solution process of RIM decomposes the network into k *Basic Network Components (BNC)*. Each BNC contains a target node v as the only one node in the target layer and v has either zero hop (A), one hop (B) or two (C) hops of in-neighbors (see Fig. 2). Here u_is are the first hop in-neighbors and w_is are the second hop in-neighbors of target layer node.

The mathematical formulation of RIM problem starts with calculating the marginal opportunity cost set of each v as:

$$\Gamma(v): \quad \text{minimize} \sum_{u\in n^{-1}(v)} x_u \tag{1}$$

subject to,

$$\sum_{u\in n^{-1}(v)} w_{uv} * x_u \geq \theta_v, \tag{2}$$

$$x_u \in \{0,1\}, w_{uv} \in (0,1] \tag{3}$$

Then the opportunity cost set $\Gamma(S)$ is constituted by combining all marginal cost sets of all nodes $v \in S$ as:

$$\Gamma(S) = \bigcup_{v\in S} \Gamma(v) \tag{4}$$

Finally, the opportunity cost $\sigma(S)$ is given by:

$$\sigma(S) = |\Gamma(S)| \tag{5}$$

Definition 1. RIM Problem: Given a social network $G(V, E)$ and a target set S, the RIM problem estimates the opportunity cost, $\sigma(S)$, which is defined by the minimum number of nodes that must be activated in-order to activate all the nodes in S. □

IV. Solution Frameworks of RIM

In this section, we have discussed the challenges and proposed R-RIM and RLT-RIM models to solve the RIM problem.

A. Challenges

The first challenge is to set the stopping criteria. Generally, single hop is considered in influence maximization method but we have employed a model which includes maximum two-hop in-neighbors to balance accuracy and estimation complexity trade-off. The second challenge is to handle three BNCs. The case BNC-A is trivial and we just offer free sample products to the target node [5] and set $\sigma(v) = |\Gamma(v)| = |\{v\}| = 1$. The case BNC-B is the basic unit of calculation and the case BNC-C is a combination of multiple instances of BNC-A and BNC-B. Thus, it is enough to design only BNC-A and BNC-B to tackle the RIM problem. The third challenge is to handle insufficient influence which happens when all the in-neighbors have not enough aggregated influence to activate the target node. We have applied small threshold values to avoid it. The final challenge is the NP-Hardness of RIM problem.

Theorem 1. *The RIM problem is NP-Hard.*

Proof. Consider the well-known Knapsack problem:

$$\text{maximize} \sum_{u=1}^{N} x_u p_u \tag{6}$$

subject to,

$$\sum_{u=1}^{N} x_u w_u \le M, \tag{7}$$

$$x_u \in \{0, 1\} \tag{8}$$

Here w_u and p_u are the weights an associated profits of N items. Now let us consider influence weights w_{uv} in RIM as item weights w_u in the knapsack problem, node's threshold θ_v as the knapsack size M and substitute the objective function of the RIM problem by $maximize\ -\sum_{u \in n^{-1}(v)} x_u$. The Knapsack problem, which is NP-Hard [14], is reduced to the RIM problem. Thus the RIM is NP Hard. □

Algorithm 1: R-RIM Model

```
Input: G(V, E), S
Result: Γ(v)
TSS = ∅ MSS = ∅;
for each v ∈ S do
    MSS = n^{-1}(v);                 /* first hop neighbors */
    for each u ∈ S do
        MSS = MSS ∪ n^{-1}(u) ;      /* 2nd hop ... */
    end
    TSS = TSS ∪ MSS;  /* Total Solution Space */
end
m = Select a number between(1, |TSS|) randomly;
Γ(S) = Select m nodes from TSS randomly;
return Γ(S);
```

Algorithm 2: MarginalCost

```
Input: G(V, E), S, v, p, q
Result: Γ(v)
Calculate the set n^{-1}(v);
active = 0;
if n^{-1}(v) = ∅ then
    t_v = 1;
    return v ;                  /* BNC-A: No incoming node */
end
inf_sum = 0, innset = n^{-1}(v), Γ(v) = ∅;
while innset ≠ ∅ do
    if inf_sum ≥ θ_v then
        active = 1, q_u = 1;
        break;
    end
    u = Select a node from 'innset' randomly;
    p_u = 1;
    inf_sum = inf_sum + w_uv;
    Γ(v) = Γ(v) ∪ u; innset = innset − u;  /* Selects
     in-neighbors for LT */
end
return Γ(v);
```

B. The R-RIM algorithm

The Random RIM (R-RIM), stated in the Alg. 1, calculates the Marginal Solution Space (MSS) set by taking first hop in-neighbors of all k target nodes. Then the Total Solution Space (TSS) is formed by taking all the members of the MSS along with the second hop in-neighbors. Finally, a random number of nodes are selected from *TSS* as the opportunity cost set, $\Gamma(S)$ and the final opportunity cost is $\sigma(S) = |\Gamma(S)|$. The complexity of the algorithm is defined by the number of in-neighbor nodes required to be explored and is given by $O(kd^2)$, where d is the maximum in-degree.

Algorithm 3: RLT-RIM Model

```
Input: G(V, E), S
Result: σ(S), Γ(S)
Γ(S) = ∅;                         /* First hop neighbors */
for each v ∈ S do
    Γ(S) = Γ(S) ∪ MarginalCost(G, S, v, y, t) by
     equation (1) to (3);          /* Equation (4) */
end
S_1 = Γ(S) ;                      /* Second hop neighbors */
for each v ∈ S_1 do
    Γ(S) = Γ(S) ∪ MarginalCost(G, S_1, v, x, y) by
     equation (1) to (3);          /* Equation (4) */
end
return σ(S) = |Γ(S)|;             /* Equation (5) */
```

C. The RLT-RIM algorithm

The RLT-RIM algorithm, as stated in the Alg. 3 and Alg. 2, is an extension of LT model. It randomly selects an in-neighbor u of a target node v and aggregate its influence weight w_{uv}. If the aggregated influence reaches to threshold θ_v, then v is activated and all the selected in-neighbors are included in marginal cost set $\Gamma(v)$. Finally, all the marginal cost sets are merged to find the opportunity cost $\sigma(S) = |\Gamma(S)|$. The best case happens when the algorithm selects only one in-neighbor node and the worst case occurs when the algorithm chooses all the in-neighbors from both the first and second hop in-neighbors. The complexities are $O(k)$ and $O(kd^2)$ respectively. In the average case, the algorithm picks up the expected number of in-neighbors with equal probability $(\frac{1}{d})$ from both the first and second hop neighbors and hence the complexity is $O(kd^2)$.

V. Performance Evaluation

In this section, we evaluate the performance of the proposed R-RIM and RLT-RIM models using two real datasets.

A. Data collection

We have collected Facebook[1] and Twitter[2] datasets in which entries are given by lists of edges. The summary of each dataset is stated in the Table I.

Table I: Dataset description

Networks	**ego-Facebook** [9]	**Twitter** [9]
Nodes	$4,039$	$81,306$
Edges	$88,234$	$1,768,149$
Average cluster coefficient	0.6055	0.5653

[1] https://snap.stanford.edu/data/egonets-Facebook.html
[2] https://snap.stanford.edu/data/egonets-Twitter.html

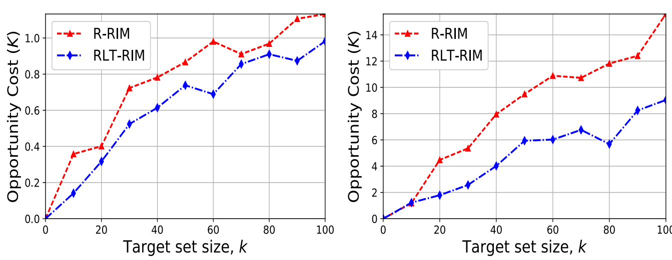
B. Simulation setup

We have employed Monte Carlo (MC) simulation [2] on both the datasets to evaluate the performance of the proposed algorithms. Each algorithm is executed 100 times and the average is taken. The target set S is generated randomly for simplicity. We have adopted *degree centrality* [2] technique to compute the influence weight w_{uv}. In the experiment, the threshold value, θ_v is assigned by applying the Heuristic Individual (HI) threshold model [15] which gives favorable values to reduce the insufficient influence effect.

C. Performance analysis

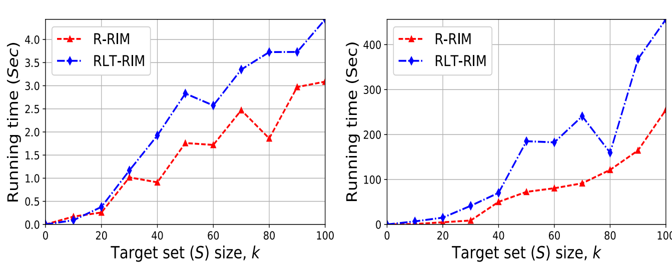
The RLT-RIM model exhibits the lower cost margin than that of R-RIM model as revealed in the Fig. 3. This is due to their working principles. The R-RIM model just selects a random number of in-neighbors as opportunity cost where as the RLT-RIM checks whether the aggregated influence weight reaches the node's threshold or not each time it picks up a new. If it reaches, the target node is activated. When the target node is activated it stops probing more in-neighbors. This makes the RLT-RIM more economical than a purely random process like R-RIM. There may be some fluctuation due to the random nature of R-RIM and it is present in the Fig. 3 (b).

On the other hand, the R-RIM model shows better running time than that of RLT-RIM model as depicted in the Fig. 4. This is due to extra calculations and comparisons involved in the RLT-RIM model. Each time a node is chosen, the algorithm checks whether the aggregated influence reaches node's threshold or not. If the target node is not activated it iteratively probes other nodes consuming more time than R-RIM model which just selects a number of nodes randomly.



(a) Facebook data. (b) Twitter data.

Figure 3: The opportunity cost (in thousands) for different k.



(a) Facebook data. (b) Twitter data.

Figure 4: The running time (in sec) for different k.

VI. Conclusion

In this research, we have introduced a novel Reverse Influence Maximization (RIM) problem which estimates the opportunity cost given by the minimum number of in-neighbors that must be activated so that all the target nodes are activated. We have simulated our proposed Random RIM (R-RIM) and Randomized Linear Threshold model (RLT-RIM) using datasets of two real and prominent social networks: Facebook and Twitter. The simulation reveals that both the models show fast running time with good opportunity cost margin.

The future scope includes improving optimality of the R-RIM and RLT-RIM models which yet they provide a feasible solution along with faster running time.

References

[1] J.-R. Lee and C.-W. Chung, "A query approach for influence maximization on specific users in social networks," *IEEE Transactions on knowledge and data engineering*, vol. 27, no. 2, pp. 340–353, 2015.

[2] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 137–146.

[3] K. Burke, "As consumer attitudes shift, so must marketing strategies," in *Twelfth ACM Conference on Electronic Commerce (EC'11)*, 2003, pp. 157–166.

[4] H. Zhang, H. Zhang, A. Kuhnle, and M. T. Thai, "Profit maximization for multiple products in online social networks," in *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*. IEEE, 2016, pp. 1–9.

[5] S. Bhagat, A. Goyal, and L. V. Lakshmanan, "Maximizing product adoption in social networks," in *Proceedings of the fifth ACM international conference on Web search and data mining*. ACM, 2012, pp. 603–612.

[6] A. Talukder, R. Kamal, A. K. Bairagi, M. G. R. Alam, S. F. Abedin, M. A. Layek, H. T. Nguyen, and C. S. Hong, "Rumors in the social network: Finding the offenders using influence maximization," in *KoreanComputer Congress (KCC)*. KCC, 2015, pp. 1214–1216.

[7] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2007, pp. 420–429.

[8] B. Ryan, *Strategic accounting for management*. Cengage Learning EMEA, 1995.

[9] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks." in *NIPS*, vol. 2012, 2012, pp. 548–56.

[10] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincon, X. Sun, Y. Wang, W. Wei, and Y. Yuan, "Influence maximization in social networks when negative opinions may emerge and propagate." in *SDM*, vol. 11. SIAM, 2011, pp. 379–390.

[11] A. Goyal, W. Lu, and L. V. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 211–220.

[12] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 199–208.

[13] D. T. Nguyen, S. Das, and M. T. Thai, "Influence maximization in multiple online social networks," in *Global Communications Conference (GLOBECOM), 2013 IEEE*. IEEE, 2013, pp. 3060–3065.

[14] E. Horowitz and S. Sahni, *Fundamentals of computer algorithms*. Computer Science Press, 1978.

[15] 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 *Korean Institute of Information Scientists and Engineers (KIISE)*. KIISE, 2016, pp. 888–890.