A Novel Approach of Viral Marketing in Social Networks

Ashis Talukder, Md Abu Layek, and Choong Seon Hong
Department of Computer Engineering, Kyung Hee University, South Korea
Email: {ashis, layek, cshong}@khu.ac.kr

Abstract

Estimating influential nodes in the social network is crucial for viral or target marketing. Influence maximization (IM) approach estimates a set of such prominent users in social network. But most of the studies have not analyzed the cost of the influence maximization problem. In this research, we formulate a novel Reverse Influence Maximization (RIM) problem for cost minimization of viral or target marketing in social network and the cost is defined by the minimum number of nodes that might be activated in order to motivate a set of target nodes. The IM gives profit analysis whereas RIM offers cost analysis and together they can provide cost-benefit-analysis (CBA). We propose two random models to solve the RIM problem as well. We also perform simulation to evaluate the performance of the algorithms using two real world datasets. The result shows that the models determine the optimized opportunity cost with fast running time margin.

1. Introduction

Information and influence are defused in the social network in *word-of-mouth* [1] effect. Finding influential users in the network is very essential for viral or target marketing. The *Influence Maximization (IM)* problem finds such a small seed set that maximizes the spread of influence in the network [1]. In IM, the seed nodes are assumed to be activated initially but most of the studies have not analyzed the cost of activating them. In this research, we introduce a novel *Reverse Influence Maximization (RIM)* problem which works in opposite order as compared to IM problem. The RIM finds the opportunity cost [2] which is the minimum number of nodes that must be activated in order to motivate a given set of target nodes. These targeted members are suppose to be influential persons or brands. For example, Lionel Messi's Facebook photo using a Sumsung cellphone can influence his millions of fans to buy the same phone. Thus, the RIM problem has great business value like the IM probelm.

We prove that the RIM problem is NP-Hard and propose Random RIM (R-RIM) and Randomized Linear Threshold RIM (RLT-RIM) models to handle the RIM problem. We also evaluate the performance of the these two models using real social network datasets.

The organization of the rest of the paper is as follows: the section II contains literature overview. The section III and IV cover problem formulation and solution processes of RIM respectively. Performance analysis is stated in the section V. Conclusion and future scope have been annotated in the section VI.

2. Literature Review

The IM problem was first introduced by [3] in 2001 in social science research. Then Kempe et al. [1] have nominal work and they have formulated two classical models named Linear threshold (LT) and Independent cascade (IC) models to maximize influence in social network.

Leskovecthe et al. propose a heuristic approximation for outbreak detection using IM [4]. Goyal et al. [5] formulate a path based algorithm that shows better result than many existing models in the scale of running time, memory utilization and seed quality. Chen et al. introduce a heuristic approach in which nodes with higher degree are selected first as seed nodes [6]. It improves the accuracy of classical models [1] and the running time of CELF model [4] simultaneously.

There are many time-bounded applications ($e.g.$ promotion of a concert) where IM can be applied. Hence time is also an important factor for influence propagation. Kim et al. [7] have conducted such a work where time is incorporated in influence calculation. The authors have introduced an IC based model that seeks good quality seeds and outperforms baseline greedy model. The authors in [8] have applied influence maximization in location based product promotion and shown their algorithms are highly effective.

But none of the above studies has addressed the problem of finding the opportunity cost or seeding cost. We prose two such model to estimate the opportunity cost of viral marketing in the social network.

3. Problem Formulation

Let $G(V, E)$ is a social network, where vertices are users and edges represents social ties among them. A user u influences another user v with probability w_{uv}. The node v is activated if the combined influence of all the active in-neighbors greater than or equal to a given threshold θ_v that is $\sum_{u \in n^{-1}(v)} w_{uv} * x_u \geq \theta_v$ [1]. Here x_u indicates whether a node u is activated ($x_u = 1$) as second hop node of any target node or not ($x_u = 0$) and same definition holds for y_u and t_u for first hop node and target layer node respectively. For a given set S, $|S| = k$, of target nodes, the RIM estimates the opportunity cost set, $\Gamma(S)$ and the opportunity cost, $\sigma(S) = |\Gamma(S)|$.

The solution process of RIM decomposes the network into k *Basic Network Components (BNC)*. Each BNC has

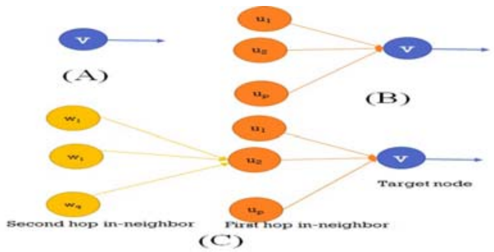


Figure 1: Basic network components of RIM

a structure that it contains a target node v as the only one node at the target layer and v has zero (A) or one (B) or two (C) hops of in-neighbors as depicted in the Fig. 1. The marginal opportunity cost $\sigma(v)$ for each target node v (for each BNC) is computed by:

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

subject to-,

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

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

Then the final opportunity cost and is given by:

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

4. Solution Frameworks of RIM

In this section we discuss the challenges of RIM problem and propose two random algorithms to handle them. One is Random RIM (R-RIM) model and the other is Randomized Linear Threshold RIM (RLT-RIM) Model.

A. Challenges

The first challenge we face is the number of predecessor hops upto which the cost calculation will be continued. Generally single predecessor hop is considered in IM problem but for better accuracy we continue estimation process upto two predecessor hops.

The decomposition process involved in the RIM solution model results in three BNCs and solving these three BNCs is the second challenge. The BNC-A is a trivial case and we set $\sigma(v) = |\Gamma(v)| = |\{v\}| = 1$. The BNC-B is the basic unit of calculation and the BNC-C is a combination of BNC-A and BNC-B. Hence designing BNC-A and BNC-B is enough to handle RIM problem. The next challenge is the NP-hardness of the RIM problem.

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

Proof. The RIM problem defined in the equations (1) to (4) can be reduced to Knapsack problem by setting node's threshold to Knapsack size, influence weights to item weights and substituting the objective function of the RIM problem by $maximize \ -\sum_{u \in n^{-1}(v)} x_u$ in the equation (1). Since Knapsack problem is a well known NP-Hard problem, RIM is also NP Hard problem. □

Algorithm 1: R-RIM Model

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

Algorithm 2: RLT-RIM Model

Input: $G(V,E), S$
Result: $\sigma(S), \Gamma(S)$
$\Gamma(S) = \emptyset$; /* First hop neighbors */
for *each* $v \in S$ **do**
  $\Gamma(S) = \Gamma(S) \cup Marginal\ Cost(v)$ by equation (1) to (3) selecting each node randomly to check node threshold θ_v;
end
$S_1 = \Gamma(S)$, $\Gamma(S) = \emptyset$; /* Second hop neighbors */
for *each* $v \in S_1$ **do**
  $\Gamma(S_1) = \Gamma(S_1) \cup Marginal\ Cost(v)$ by equation (1) to (3) selecting each node randomly to check node threshold θ_v;
end
$return\ \sigma(S) = |\Gamma(S)|$;

B. The R-RIM algorithm

The Random RIM (R-RIM) model, stated in the Alg. 1, calculates 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 second hop in-neighbors of all MSS members. Finally, a random number of nodes are selected from *TSS* as the opportunity cost.

We define the complexity of the algorithm in terms of the number of neighbors needed to be explored and the complexity of the R-RIM model is $O(kd^2)$, where d = maximum number of in-degree in the network.

C. The RLT-RIM algorithm

The RLT-RIM algorithm is stated in the *Alg. 2* which iteratively calculates the marginal cost of each target node by the equations (1) to (4). The algorithm is based on classical Linear Threshold (LT) model and at every step, it picks up an in-neighbor randomly and check whether the total influence weight has exceeded the target node's threshold or not. If the influence weight exceeds the target node's threshold, then the target node is activated.

The complexity of the RLT-RIM deserves details description. In the best case the algorithm selects only one in-neighbor node from both the first hop and second hop neighbors to activate the associated target nodes and is given by $O(k)$. The worst case occurs when the algorithm selects all the in-neighbors to do the same and is expressed by $O(kd^2)$. Finally, the average case happens when the algorithm picks up expected number of in-neighbors with equal probability $(\frac{1}{d})$ from each of first and second hop neighbors in order to activate the target nodes and is stated as $O(kd^2)$.

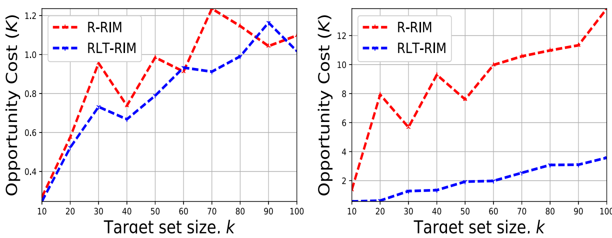
5. Performance Evaluation

We evaluate the performance of R-RIM and RLT-RIM models by taking the average values of parameters of running Monte Carlo (MC) simulation [1] for 100 times on both the datasets: Facebook[1] and Twitter[2] which are stated in the Table I. We take Heuristic Individual (HI) threshold model [9] for RLT-RIM model to generate node thresholds.

Table I: Dataset description

Networks	**ego-Facebook** [10]	**Twitter** [10]
Nodes	$4,039$	$81,306$
Edges	$88,234$	$1,768,149$
Average cluster coefficient	0.6055	0.5653

The R-RIM model shows better running time than that of RLT-RIM model as it just randomly selects a number of in neighbors as depicted in the Fig. 3. But on the other hand, in the Fig 2, the RLT-RIM model exhibits the lower cost margin than that of R-RIM model science, each time the RLT-RIM model picks up a node, it checks whether the exaggerated influence reach to the threshold or not. When the target node is activated it stops probing more in-neighbors. For both the datasets, some fluctuations are present in the result due to the inherent degree distribution and random nature of the algorithms.

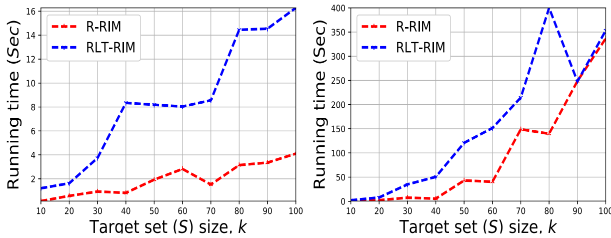


(a) Facebook data. (b) Twitter data.

Figure 2: The opportunity cost for different k values.

6. Conclusion

In this work, we introduce a novel Reverse Influence Maximization (RIM) problem and proposed two solution models,



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

Figure 3: The running time (in sec) for different k values.

Random RIM (R-RIM) and Randomized Linear Threshold model (RLT-RIM) to estimate the optimized opportunity cost for target marketing in social network. Both the models show good running time with optimized opportunity cost margin.

The R-RIM and RLT-RIM models can not always ensure optimality for its random nature but yet provide a feasible solution with faster running time. This has been left as future scope of this work.

***Acknowledgements*—This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (B0190-16-2013, Development of Access Technology Agnostic Next-Generation Networking Technology for Wired-Wireless Converged Networks) *Dr. CS Hong is the corresponding author.**

References

[1] 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.

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

[3] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[4] 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.

[5] 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.

[6] 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.

[7] J. Kim, W. Lee, and H. Yu, "Ct-ic: Continuously activated and time-restricted independent cascade model for viral marketing," *Knowledge-Based Systems*, vol. 62, pp. 57–68, 2014.

[8] T. Zhou, J. Cao, B. Liu, S. Xu, Z. Zhu, and J. Luo, "Location-based influence maximization in social networks," in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 2015, pp. 1211–1220.

[9] A. Talukder, M. G. R. Alam, A. K. Bairagi, S. F. Abedin, H. T. Nguye, 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.

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

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