Scalable Influence-Aware Profit Maximization Over Livestreaming Marketing Network

Hao Du[1]

[1]KuiMian42@gmail.com

[1]China University of Petroleum - Beijing at Karamay, Karamay, China

Abstract. Profit maximization (PM), with the purpose to select the appropriate set of initial seed users to maximize the effectiveness of diffusion, has become the focus of Social Network Analysis with broad prospects of applications such as social opinion propagation and Internet marketing. Under the condition of ensuring great performance, the existing PM models are faced with the challenge of time complexity and universality because they take too long to execute and their working conditions come-= with harsh restrictions. In this paper, we propose a new algorithm named CirclePrune (CP) which optimizes the runtime in large-scale network and loosens the constraints by warming up, and apply it to the scenario of livestreaming marketing. Experimental results confirm the effectiveness and efficiency for the CP algorithm.

Key words: Profit maximization, loose-constrained, warm up

1 INTRODUCTION

Social platforms like Twitter and Weibo have numerous active users whose every move has brought vast amounts of information and commercial value over the last few decades. And this is the main factor driving the booming development of Social Network Analysis [1]. Nowadays, many people willingly share their information with the world by releasing friends' circles, trends in moments and other ways, which has triggered the upsurge of research on diffusion through social networks. Hence, the profit maximization (PM) algorithm gradually takes to the stage with considerable economic benefits.

PM algorithm is a propagation algorithm that considers both influence and cost. The diffusion of a realistic network is often associated with cost before spread which might place severe limitations, especially in economic activity. With the advance of the PM algorithm, plentiful models studied in various domains are used to simulate influence propagated through a social network, including the popularization of the excellent public reputation of the new product and recommending new friends [2]. In addition, studies on preventing rumor spreading [3] and viral marketing [4] strategies designed for boosting sales also prove the importance of PM for information diffusion research and social network development.

The effect of the existing propagation model heavily depends on the cost of the initial seed users. And the efficiency of acquiring the target node set relies on the method chosen and the

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322630

probability of propagation. Furthermore, to improve the model's universality, we should pay close attention to how to make the constraints of the model weaker and easier to reach.

In this paper, we firstly revisit two greedy algorithms based on Independent Cascade (IC) model [5]: Simple Greedy (SG) algorithm [6] and Double Greedy (DG) algorithm [7]. The latter significantly reduces runtime and has weaker constraints at the expense of performance compared to the former one. Based on the conditions of the DG algorithm, we find some scenarios like the marginal nodes of the community with expensive costs could not meet the conditions of the DG algorithm. Therefore, we propose the CirclePrune (CP) algorithm, which warms up for the DG algorithm and loosens the constraints again while maintaining the same performance as the DG algorithm. Then, we treat the dataset obtained from the host and fans in live streaming marketing on Bilibili. And we simulate algorithms in different settings with respect to replacing the probability of spread p with two different probabilities in diffusion, which is closer to the real world because the influence between fans is not necessarily the same as fans being directly influenced by the celebrity. Experimental results confirm the effectiveness and efficiency of the CP algorithm.

The rest is organized as follows. Section II presents the basic knowledge regarding graph theory and the IC model. Section III introduces three algorithms with their efficiency and constraints in turn. The next part compares the CP algorithm with other algorithms on the dataset we collected. In section V we summarize this paper.

2 BASIC KNOWLEDGE

2.1 Graph Theory

The graph is an abstract representation with nodes and a set of edges. In this structure, nodes represent entities, and edges represent the relationships between described entities. We specify the type of graph based on whether the edges have weights and directions. Here we focus on unweighted undirected graphs and give an example, Graph G shown in Fig. 1. The degree is an important concept of graph structure, and a node's degree describes the number of edges connected to it. For instance, node 3 in G connects with five edges, so its degree is 5. The shortest path focuses on the graph structure, and many algorithms are derived. It describes the shortest distance from one node to the other node. In G, the shortest path between node 2 and node 4 is {2, 3, 4} instead of {2, 1, 3, 4} or others.

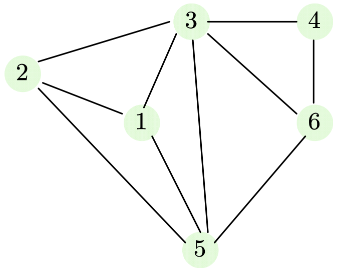

Figure 1. Graph GIndependent Cascade Model

Centrality, which is also a measure of the importance of a node in a graph, defines as a node that is considered significant if many other significant nodes are connected to it. The standard measures include the degree centrality, the eigenvector centrality, the betweenness centrality, and the closeness centrality. Here are two of them that we apply to the experiment. The degree centrality is the node's degree, and the definition of closeness centrality is the average length of all shortest paths via the node.

There is two widely used information diffusion mode named Threshold Model of Diffusion [8] and Cascade Model of Diffusion [9]. The simplest and most popular form of the Cascade Model is Independent Cascade (IC) Model. IC model is a stochastic information diffusion model where the influence spreads the network through Cascade. The node in the IC model has two states: active and inactive. It means whether the information diffusion influenced the node.

The IC model focuses on the strategy [6] of selecting seed users, which has a substantial impact on the influence of the model, and we conduct a test in Section IV. We select the initial seed users to turn into the active state at the beginning of diffusion. The active node set propagates with a certain probability p to active node neighbors and lets the newly activated nodes spread in the next round. Next, repeat the process until no longer getting the newly activated nodes.

3 PROFIT MAXIMIZATION ALGORITHM

In this section, we first revisit two PM algorithms and propose a new optimization algorithm by warming up to loosen the constraints of the DG algorithm. It also can reduce runtime in large-scale networks. Each algorithm significantly improves on the previous one in some aspects, such as reducing the time complexity or holding for looser constraints.

3.1 Simple Greedy Algorithm

The first algorithm, named Simple Greedy (SG), mainly focuses on the strategy of selecting seed users. It proved that seed users selected by the SG algorithm get the best score in the experiment in the next part. Assuming that we ignore the cost of the initial seed users, given the factor that influences spread function under the IC model is submodular and monotone. For any two node sets $S \subseteq T$ and any node $v \notin T$, the submodular function $\sigma(\cdot)$ satisfies

$$\sigma(S \cup v) - \sigma(S) \geq \sigma(T \cup v) - \sigma(T) \tag{1}$$

Because of these two properties and formula 1, we can establish a worst-case, lower bound on the maximum number according to the theory of the greedy algorithm. The conclusion is formula 2 presented in this paper [6].

$$\sigma(S_k) \geq \left(1 - \frac{1}{e}\right)\sigma(S^*) \tag{2}$$

In the above formula, $\sigma(\cdot)$ stands for the influence function and it also means the profit based on the previous assumption. Then, S_k is a set of initial seed users which is composed by k nodes. And S^* represent the most profitable initial node set.

The process of SG algorithm is presented.

The time complexity is $O(V^2M)$ and it's unbearable.

Algorithm 1 *Simple Greedy*

```
start with $S \leftarrow \emptyset$ ;
while the size of $S < k$ do
    find $t \leftarrow \arg\max_{v \in V \setminus S} \{\sigma(S \cup v) - \sigma(S)\}$ ;
    $S \leftarrow S \cup \{t\}$ ;
return $S$
```

3.2 Double Greedy Algorithm

The constraint conditions applied to the SG algorithm are exceedingly harsh as spread cost causes the IC model to no longer obtains the property of monotone. Under the circumstances, the SG algorithm can perform arbitrarily worse than the optimal solution. There are two versions: Deterministic Double Greedy Algorithm and Random Double Greedy Algorithm. The emergence of the Double Greedy (DG) Algorithm has notably improved the efficiency and effectiveness of the influence propagation. The DG algorithm, which gets rid of the monotone condition, has a diversity of application scenarios and satisfy the requirement that the profit of V is positive, such as formula 3 below:

$$\varphi(V) = \sigma(V) - c(V) \geq 0 \tag{3}$$

The equation indicates that the profit function $\varphi(\cdot)$ equals the influence function $\sigma(\cdot)$ minus the cost function $c(\cdot)$.

Similarly, the DG algorithm with formula 4 and formula 5 gives a specific boundary:

$$\varphi\big((S^* \cup S_0) \cap T_0\big) + \varphi(S_0) + \varphi(T_0) \leq 3 \cdot \varphi(S_d) \tag{4}$$

$$\varphi\big((S^* \cup S_0) \cap T_0\big) + \varphi(S_0) + \varphi(T_0) \leq 2 \cdot E[\varphi(S_R)] \tag{5}$$

Those two nodes set S_0 and T_0 stand for the initial set before executing the DG algorithm. S_d is the result of the Deterministic Double Greedy Algorithm and S_R is the result of the Random Double Greedy Algorithm.

The pseudocode for the DG algorithm is as follows. For the random version, we change the condition of line 5 to $\mathrm{U}(0,1) \leq \frac{r^+}{r^+ + r^-}$, where $\mathrm{U}(0,1)$ is a uniformly distributed number between 0 and 1, and $\frac{r^+}{r^+ + r^-} = 1$ if $r^+ + r^- = 0$.

Algorithm 2 *Deterministic Double Greedy*

```
1 start with $S \leftarrow \emptyset$ , $T \leftarrow V$ ;
2 for $u \in V$ do
3     $r^+ \leftarrow \varphi(S \cup \{u\}) - \varphi(S)$ ;
4     $r^- \leftarrow \varphi(T \setminus \{u\}) - \varphi(T)$ ;
5     if $r^+ \geq r^+$ then
6         $S \leftarrow S \cup \{u\}$ ;
7     else
8         $T \leftarrow T \setminus \{u\}$ ;
9 return $S (= T)$
```

It's evident that the time complexity of the DG algorithm is $O(VM)$ and therefore, we can draw a conclusion that the DG algorithm remarkably lowers the runtime of the diffusion process.

3.3 CirclePrune

Based on the observation of the conditions of the DG algorithm, we found that some scenarios, such as the marginal nodes of the community with expensive costs, could not meet the conditions of the DG algorithm. It's imperative that we need to find an algorithm with weaker constraints to accommodate. Hence, we purpose the CP algorithm, and it's just a warm-up step for the DG algorithm. By utilizing the prune measure, we can gain an extremely tight lattice regarded as the initial boundary at the beginning of the DG algorithm. The lattice retains all global maximizers for the profit function, for instance, for all where $A^* \leq S^* \leq B^*$ for all S^* where $\varphi(S^*) = max_{S \in V}\, \varphi(S)$.

Proof: For any node set $S \subseteq T \setminus \{v\}$, according to submodularity of $\varphi(*)$ and the property of monotone ascending, we can get $\varphi(S \cup \{v\}) - \varphi(S) \geq \varphi(V) - \varphi(V \setminus \{v\}) \geq 0$. Because $S \cup \{v\}$ always produces higher profit than S, so v must be selected in the initial seeds, and we will get a minimal set named $A_1 \subseteq S^*$. Similarly, we can obtain the maximizing set named B, which contains S^*. The lattice $L_1 = [A_1, B_1]$ is useful for warm-starts, and we can prune L_1 even further by using a circle strategy.

There is a corollary that we can get the ability to ensure a same profit baseline which is identical to the single DG algorithm and utilize this algorithm in the weaker restrictions by means of warming up. Instead of guaranteeing $\varphi(V) \geq 0$ before diffusing, we can achieve the goal only via a less constrained situation like formula 6:

$$\varphi(A^*) + \varphi(B^*) \geq 0 \tag{6}$$

The simple procedure of algorithm 3 is shown below:

It's important and notable that the runtime of the prune process might be slightly longer than the single DG algorithm but it's verified implementing the warm-up method on a large-scale network will speeds up the course of spread because each circulation in the algorithm 3 is quicker than one in the DG algorithm, though the exact number of iterations by prune is bigger.

Algorithm 3 *CirclePrune*

```
start with $t = 0,\ A_0 \leftarrow \emptyset\ ,\ B_0 \leftarrow V$ ;
repeat
    for $v \in B_t$ do:
        $A_{t+1} \leftarrow \{v : \varphi(B_t) - \varphi(B_t \setminus \{v\}) > 0\}$ ;
    for $v \in V \setminus A_t$ do:
        $B_{t+1} \leftarrow \{v : \varphi(A_t + \{v\}) - \varphi(A_t) \geq 0\}$ ;
    $t \leftarrow t + 1$ ;
until $A_t = A_{t-1},\ B_t = B_{t-1}$ ;
return $A_t$ and $B_t$
```

4 EXPERIMENT

In this section, we firstly analyze the benefits of the IC models with a varying number of seeds under the four strategies and compare changes between the fore-and-aft models on three datasets. Hence, we show the evolution of the lattice of algorithm 3 in our synthetic dataset. Finally, based on the previous theoretical knowledge, we compare various algorithms in terms of efficiency and runtime in a more realistic situation. Precisely, for the difference from the hypothesis of the IC model that fixing the probability of propagation is p, we consider the influence of the initial seed users on the others might be unequal to the influence of the affected nodes on the nodes to be spread. For example, we usually select the Internet celebrity as the initial seed users because they are influential and have numerous fans. Based on the assumption above, the probability of those celebrities spreading to their fans denotes as p_1 and the probability of diffusion among the fans stands for p_2.

4.1 Expriment Setup

The strategies of selecting seed users: There are three approaches to obtaining the initial seed users besides the SG algorithm. Naturally, we gain the seeds by choosing the ones with the most significant degree. And selecting by the closeness centrality is an interesting plan. Last, we add a random version as a contrast.

Demonstrate the pruning process: In this part, we keep track of the changes of the lattice and show how it's compressed in algorithm.

Compared algorithm: Three algorithms compared on the captured dataset with different probability of diffusion and we set the cost of the initial seed users as its one tenth degree. The values for parameters of those algorithms are as follow. We fix the probability of p_1 is 0.5 which means the influence probability from the celebrity to fans and the value of p_2 is equal to 0.3 or 0.7 which presents the spread probability between the fans. Finally, the cost is node's one tenth degree. The experimental results are shown in Figure 2 and Figure 3.

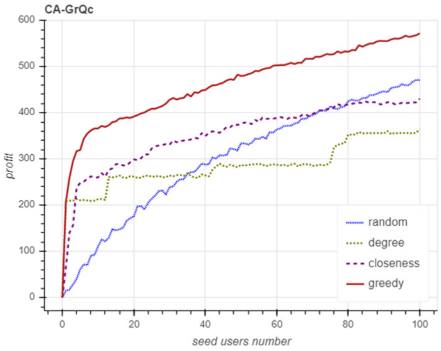

Figure 2. the influence with different seed users on CA-GrQc

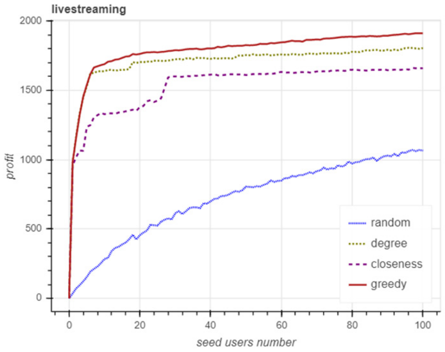

Figure 3. the influence with different seed users on livestreaming

Dataset: We use two existing datasets and the new dataset gathered from the livestreaming marketing platform. The first datasets named "CA-GrQc" from the Stanford Large Network Datasets Collection [10]. And the last anonymized dataset named "livestreaming" was got from Biliiili's studio and the reading quantity of recommended product.

4.2 Performance analysis

We plot a bar chart to show the pruning process in Fig. 4. In this chart, we can see that set A and set B to converge very quickly and have similar lengths. The runtime of executing the DG algorithm after the warm-up is negligible compared to the runtime of pruning.

The consequence of diffusion by those algorithms with constant p_1 and different p_2 on dataset 2 is shown in Fig. 5 and Fig. 6. We can find that the SG algorithm still shows the best and the others are finely balanced. And The performance differences of these algorithms simulated on this dataset are so close.

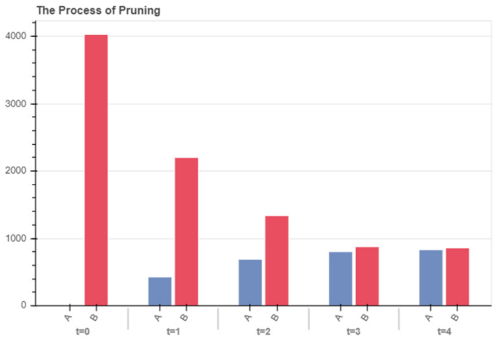

Figure 4. the process of pruning on livestreaming

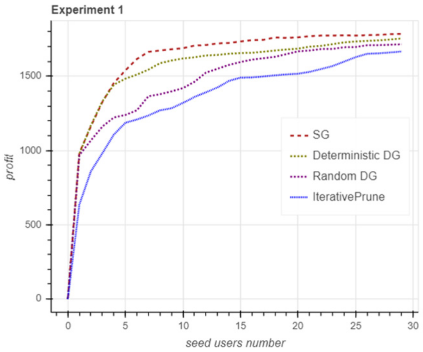

Figure 5. $p_1 = 0.5$, $p_2 = 0.7$

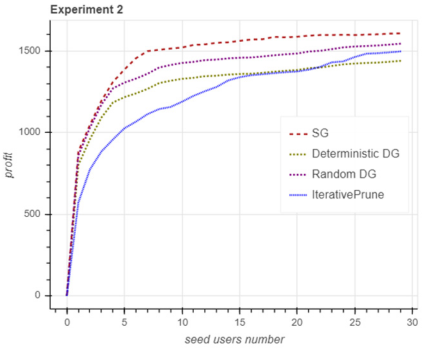

Figure 6. $p_1 = 0.5$, $p_2 = 0.7$

Each algorithm in Experiment 1 corresponds to the algorithm displayed in Fig. 7. The DG Algorithm and the CP algorithm save a lot of running times compared with the SG algorithm, and in Experiment 2 the result is similar. According to Fig. 8, we discover that the pruning round of $t = 1$ and $t = 2$ and takes much less time than the DG algorithm. For the CP algorithm, the former round has a larger range to traverse. So we can verify that the CP algorithm plays a role in speeding up the execution of large-scale network.

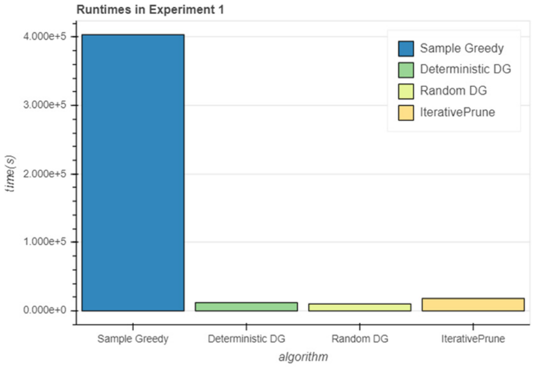

Figure 7. Runtimes in Experiment 1

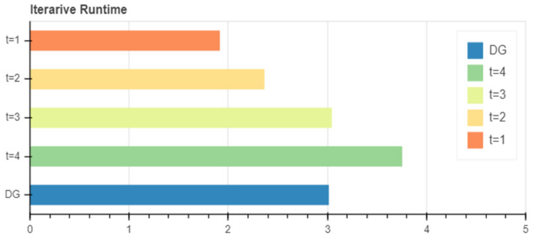

Figure 8. the runtime of each round in the CP algorithm

5 CONCLUSION

This paper aims at a new algorithm named CirclePrune for warm-up before executing the Double Greedy algorithm. We conclude that the algorithm weakens the constraints of the Double Greedy algorithm, and it also saves runtimes in a large-scale network. We collect a new dataset from the live streaming marketing on the Bilibili platform and put it into practice. Finally, we treat the CirclePrune algorithm compared to the other algorithms with respect to the profit of diffusion, the restrictions and runtimes.

REFERENCES

[1] F. Bonchi, "Influence propagation in social networks: A data mining perspective", IEEE Intell. Informat. Bull., vol. 12, no. 1, pp. 8-16, 2011.

[2] H. Zheng and J. Wu, “Friend recommendation in online social networks: Perspective of social influence maximization,” in International Conference on Computer Communication and Networks, Vancouver, Canada, pp. 1–9, 2017.

[3] G. Tong, W. Wu, L. Guo, D. Li, C. Liu, B. Liu, and D. Du, “An efficient randomized algorithm for rumor blocking in online social networks,” in Proc. of IEEE INFOCOM, Atlanta, GA, USA, pp. 1–9, 2017.

[4] J. Tang, X. Tang, and J. Yuan, “Profit maximization for viral marketing in online social networks,” in IEEE Internaltional Conference on Network Protocols, Singapore, 2016.

[5] Jacob Goldenberg, B. Libai, E. Muller: “Talk of the network: A complex systems look at the underlying process of word-of-mouth”, Marketing Letters, pp. 211—223, 2001.

[6] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In Proc. of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2003.

[7] Buchbinder, Niv, et al. "A tight linear time (1/2)-approximation for unconstrained submodular maximization." SIAM Journal on Computing 44.5 (2015): 1384-1402. M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[8] Sankar K. Pal, Suman Kundu, C. A. Murthy: “Centrality Measures, Upper Bound, and Influence Maximization in Large Scale Directed Social Networks”, Fundamenta Informaticae.

[9] Everett M. Rogers: Diffusion of Innovations. Free Press, 2003

[10] J. Leskovec, J. Kleinberg and C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007