Maximizing the Speed of Influence in Social Networks

Yubo Wang and Robert Chun
Department of Computer Science, San Jose State University, San Jose, California, US

Abstract—Influence maximization is the study of seed-node selection in a social network in order to achieve the maximized number of influenced nodes. Previous studies focused on three areas, i.e., designing propagation models, improving seed-node selection algorithms and exploiting the structure of social networks. However, most of these studies ignored the time constraint in influence propagation. Here, I studied how to maximize the speed of influence propagation. I extended the classic Independent Cascade model to a Continuous Dynamic Extended Independent Cascade (CDE-IC) model. In addition, I proposed a novel heuristic algorithm and evaluated the algorithm using two large academic collaboration data sets. The new algorithm is 10%-20% faster in influence propagation than previous classic heuristic algorithms on the CDE-IC model. Furthermore, I gave solutions to calculate propagation probability between adjacent nodes by exploiting the structure of social network.

Keywords: Social network, Influence speed maximization, Diffusion model, heuristic algorithm

1. INTRODUCTION
Social networks provide a great opportunity to promote new products or ideas because of the large number of users and the high frequency of communication. In viral marketing strategy, a company invites some initial users, i.e. the seed nodes, to try its new products or technologies and hope that these initial users would give a positive feedback. By the power of word-of-mouth, these users may affect their neighbors in a social network. These affected neighbors may subsequently affect their own neighbors. Therefore, in a fashion similar to viral infection, the influence propagates in the social network. The challenge in viral marketing strategy is how to select the seed nodes to maximize the number of affected nodes.

The large scale of social networks and their complicated structures made it challenging to select the right seed nodes. Influence maximization was first proposed as an algorithm problem by Domingos and Richardson in a study of viral marketing[1, 2]. Kempe, Kleinberg and Tardos provided a foundation to solve this problem[3]. They proved that the problem of optimization in seed-node selection is NP-hard. They also presented the first provable approximation solution, which is within 63% (1-1/e) of optimal. One big drawback in this solution is low efficiency. Even finding a small seed-node set in a moderately large network (e.g. 15000 nodes) would take days to finish. Several following studies have been carried to improve the efficiency of seed-node selection algorithms. Leskovec, Krause and Guestrin proposed a nearly optimal algorithm called Cost Effective Lazy Forward (CELF) algorithm[4], which was 700 times faster. Goyal, Lu and Lakshmanan proposed an algorithm called CELF++, which was 35%-55% faster than CELF[5].

Chen, Wang and Yang[6] tackled the efficiency issue of seed-node selection by improving heuristics methods. Their new heuristics method achieved a nearly matched result comparing to greedy algorithms, but with a running time of more than six orders of magnitude faster.

All the previous studies focused on the space maximization of influence, i.e. how to maximize influence propagation in a social network without time constraint. In this paper, I will study the speed maximization of influence propagation, i.e. how to maximize influence propagation in a social network in a given time frame.

Here, I showed how I extended a classic IC model to Continuous Dynamic Extended IC (CDE-IC) model. In addition, I proposed a novel heuristic seed-node selection algorithm. I then gave formulas to calculate propagation probability between adjacent nodes. In order to test my new algorithm, I compared this algorithm with three other most popular heuristic algorithms in two data sets. The result showed that even a small modification to the existing algorithms could lead to a big boost to the qualification of seed nodes selection.

2. RELATED WORK
Studies on influence maximization in social networks are mainly focused on three areas, designing propagation models, improving seed-node selection algorithms and exploiting social networks structure.

2.1 Propagation Models
The two basic models of social networks are Linear Threshold (LT) model and Independent Cascade (IC) model. Many other models extended from these two basic models under different conditions.

2.1.1 Linear Threshold (LT) Model
LT Model was proposed by Granovetter and Schelling to simulate influence propagation in social networks [7, 8]. In LT model, each node u has a threshold $\theta_u$ ($0<\theta_u<1$) that defines the minimum requirement of its active neighbor nodes set. When the total weights of all active neighbor nodes are greater than $\theta_u$, an inactive node u switches to an active status.
Each node has the same threshold value of 0.3. The weight between any two nodes is labeled at the edge. At time 0, node A is selected as the seed node with an active status. Node A will try to propagate influence to its neighbor nodes of Node B and Node D. Edge AB has a weight value of 0.4 and edge AD has a weight value of 0.1. In this condition, only node B satisfies the condition of switching from an inactive status to an active status as the total weight of active neighbor nodes of Node B is greater than the threshold of 0.3. Therefore, Node B turns into an active status at time 1. The total weights of active neighbors of node D are 0.1 and node D remains inactive at time 1. The following steps are similar until there are no more nodes that satisfy the condition to switch and the process of influence propagation stops.

There are two interesting observations in this propagation process. First, although some people will not accept new technologies at first, they are likely to change their minds as more of their neighbors accept the new technologies. These people are represented by node D in the example. Secondly, although the weight between node C and node F is higher than the threshold, they have no chance to switch to active status. Their active neighbor nodes are not powerful enough to propagate the influence to them. Node C and node F represent the users in social networks who are eager to accept new products. However, because inappropriate seed nodes are selected, the propagation process fails to discover these users. We have to carefully consider this during the seed nodes selection phase.

2.1.2 Independent Cascade Model

The Independent Cascade (IC) model was proposed by Goldenberg, Libai, and Muller[9]. In IC model, the process of influence propagation in social networks can be illustrated as the following steps. An initial seed-node set is selected at time 0. If node A becomes active at time t, it will try to affect its inactive neighbors at time t+1 with a probability of p. Node A will only have one chance to affect its neighbors, and will not try again whether it succeed or not. If an inactive node is connected with more than one newly active node, these newly active nodes will try to influence the inactive node in a random sequence. In addition, the result of influence propagation between two nodes is not affected by actions of other nodes.

The process of influence propagation In IC model is shown in Figure 2 and Figure 3, each representing an outcome from a single run. In both scenarios, node A is selected as the seed node. However, the end results are different because of the randomness in influence propagation in IC model. The influence reaches to 4 other nodes (D, E, F and H) in one run (Figure 2) and 5 other nodes (B, C, E, F and H) in the other run (Figure 3). Therefore, the experiments need to run multiple times in IC model to obtain an accurate estimation of influence propagation with a specific seed node selection.

2.1.3 Extensional IC Model

Wang, Qian, and Lu proposed an Extensional Independent Cascade (EIC) model[10]. In the EIC model, the influence propagation between two adjacent nodes is divided into two phases involving a spreading phase from an active node and an adopting phase from the inactive node. In the spreading phase, an active node decides whether to spread the influence to its neighbor nodes based on a spreading probability of ps. If the active node decides to spread out the influence to its neighbor nodes, the adopting phase is similar as in the original propagation process in classic IC model. In the adopting phase, each inactive node decides independently whether to adopt the influence based on an adopting probability of pa. The EIC model added one more step in the process of influence propagation. The new propagation probability (pp) in EIC model is the product of probability in spreading phase (ps) and probability in the adopting phase (pa), as in formula 1 $pp = ps \times pa$.

2.2 Seed-node Selection Algorithms

Seed-node selection algorithms play a key role in influence maximization. There are two metrics to measure a seed node selection algorithm, efficiency and quality.
2.2.1 Hill-Climbing Algorithm
Kempe et al.[3] used a greedy algorithm named Hill-Climbing algorithm. They proved that the Hill-Climbing algorithm guarantees to achieve an approximation solution with a factor \((1- 1/e – \varepsilon)\) to the optimal solution in both the LT Model and the IC Model. Here \(e\) is the natural logarithm base and \(\varepsilon\) is any small positive real number. This algorithm is based on the theory of submodular functions[11]. In order to obtain an accurate estimation, Kempe et al. run Monte Carlo simulation sufficient number of times (20,000) for selecting each seed node. Although greedy algorithm guarantees the quality of seed nodes selection, it is not efficient enough for large-scale social networks. Many later studies tried to improve the efficiency of this algorithm, including CELF Algorithm and CELF++ Algorithm.

2.2.2 CELF and CELF++ Algorithm
The biggest drawback of the Hill-Climbing greedy algorithm is low efficiency. Leskovec et al. tried to solve this problem using an improved greedy algorithm named Cost-Effective Lazy Forward (CELF)[4]. Their experiment showed that they would get a near optimal solution while being 700 times faster than the Hill-Climbing greedy algorithm. Further effort of reducing the running time of the greedy algorithm was carried by Goyal et al.[5]. They proposed an improved CELF algorithm called “CELF++” that further decreases the running time by 35%-50% compared to the CELF algorithm. Despite those improvements, use of greedy algorithms in large data sets is still limited by the relatively low efficiency.

2.2.3 Heuristic Algorithm
In contrast to greedy algorithms, heuristic algorithms may not provide the best result, but they can obtain an acceptable result in much less time. Two widely used heuristic algorithms are Degree-Centered and Distance-Centered.

In Degree-Centered algorithm, the nodes that have a large number of connections in a social network, i.e. the high degree nodes, are deemed as influential nodes. The more out-neighbors a node has, the more influence it is believed to have in a social network. This is an intuitive assumption. However, a known phenomenon in social network is that high degree nodes tend to connect to each other. Therefore, if only high degree nodes are selected as seed nodes, these seed nodes will have a large overlap with each other. The overlapped nodes will not bring any additional value to the set of seed nodes to maximize influence propagation. In Distance-centered algorithm, influential nodes are the nodes that have a smaller average distance to other nodes.

Although Degree-Centered algorithm is better than other heuristic algorithms[3], it is still not as good as greedy algorithms. In general, heuristic algorithms were not studied extensively in research field because of the low expectation of quality. However, Chen et al. proposed an improved heuristic algorithm in IC Model[6]. This improved heuristic algorithm performed comparably to greedy algorithms. Importantly, this algorithm significantly reduced the time in the seed-node selection with six orders of magnitude.

In this new heuristic algorithm, Chen et al. introduced the concept of “discount” to the degree of a node. The logic is that when a node is selected as a seed node, its neighbor nodes will become less influential to the social network. Therefore, there should be a discount on the neighbor nodes. They proposed two methods to discount the degree of a node. The first method is called “Single Discount”, in which the degree of all neighbor nodes of a selected seed node is reduced by one. The second method is more complicated in that the effect of each newly activated node to its inactive neighbor nodes will be calculated individually.

3. SOLUTION FRAMEWORK
In this section, I study how to maximize the speed of influence propagation in social networks. I extended the classic IC model to a Continuous Dynamic Extended Independent Cascade (CDE-IC) model. In addition, I proposed a novel heuristic algorithm comparing with previous classic heuristic algorithms on the CDE-IC model in two large academic collaboration data sets. Furthermore, I discuss how to decide the parameters in the propagation model by exploiting the structure of social networks.

3.1 CDE-IC Model
The Extensional Independent Cascade (EIC) model extends the process of influence propagation from one phase in classic IC model to two phases involving a spreading phase and an adopting phase[10]. The EIC model with two phases in propagation process is a good extension to IC model, and it is more close to reality in social networks. However, The EIC model has the same drawbacks as the classic IC model, in that the propagation process is one-time and static. One-time refers to the assumption that a newly activated node will only try to propagate the influence to its neighbors once. If a node turns into active status in step t, it will try to propagate its influence to its inactive neighbor nodes in step t+1. Static in EIC model means that the probability of propagation does not change with time.

In this paper, I propose a new improvement on the EIC model to take into consideration of continuous influence, and also the dynamic nature of influence propagation. In this Continuous Dynamic Extended IC (CDE-IC) Model, an active node will keep propagating influence to its neighbor nodes until there are no more inactive nodes. Furthermore, the probability of propagation between two nodes will change with time. The process of influence propagation in CDE-IC model is shown in Figure 4. Suppose all edges in Figure 4 have a probability of \(\frac{1}{2}\). At time 0, node A is selected as a seed node. At time 1, node B turns active under node A’s influence. At time 2, node B propagates influence to its neighbor node C and node E. At time 3, although node A fails to influence node D at time 0, it gets another opportunity and succeeds. Node D becomes active, so does node F. The influence process stops at time 3.
The key difference between CE-IC Model with classic IC Model or EIC model is that an active node will have multiple changes to propagate its influence. If it fails the first time, it still has a second or a third chance to propagate. In CE-IC model, all activated nodes need to be considered in each step.

3.2 Zero-Discount Algorithm

The Hill-Climbing algorithm and its improvements can provide guaranteed approximation solution to optimal seed node selection, but they are not efficient. Traditional heuristic algorithms such as Degree-Centered algorithms and Distance-Centered algorithms are efficient in seed nodes selection, but they do not have comparable results to greedy algorithm. The quality of Degree-Centered algorithm is lower than that of Hill-Climbing algorithms because there is a high degree of overlapping between high degree nodes. If we only select nodes with the highest degree into the seed nodes set, they will greatly weaken each other in the process of influence propagation.

To minimize overlapping between high degree nodes, I propose Zero-Discount algorithm. In this algorithm, if a node is selected as a seed node, the degree of all of its out-neighbor nodes is set to zero. In CDE-IC model, an active node will continuously propagate its influence to its neighbor nodes. If the probability of propagation between two nodes is \( p \), it only needs \( 1/p \) on average attempts to succeed. Therefore, we can safely remove all out-neighbor nodes of a seed node without impairing performance significantly. At the same time, a more evenly distributed seed node set increases the chance of influence propagation in social networks.

3.3 Estimation of Propagation Probability

The propagation probability can be set randomly or given by data mining technology. However, it will be either inaccurate or costly. Here, an estimated probability value is assigned to each edge by analyzing the characters of social networks. Influence sources are categorized into three classes, Star Effect, Peer Pressure and Social Trend.

In CDE-IC model, the propagation probability is the product of spreading probability and adopting probability. A different spreading probability is given to each node according to its degree in social networks. High degree nodes will have higher spreading probability than low degree nodes. A maximum spreading probability \( p_{\text{max}} \) is given to the node with the highest degree (\( p_{\text{max}} \leq 1 \)) and an isolated node will have a spreading probability of 0. The other nodes will be assigned a probability value based on a linear function. The spreading probability is decided as the following formula 2 by out degrees.

\[
p_{\text{s}} = p_{\text{min}} + (p_{\text{max}} - p_{\text{min}}) \frac{(\text{Degree(node)} - \text{Degree}_{\text{min}})}{(\text{Degree}_{\text{max}} - \text{Degree}_{\text{min}})}
\]

Adopting probability is estimated according to the source of influence in the following three steps. Firstly, check all in-neighbor nodes of a node. Secondly, calculate how much energy each in-neighbor node spends on it. Thirdly, normalize the total energy. In addition, it has been demonstrated that the probability of influence shows an exponential decay behavior[13]. To model this decay behavior, I introduce a time factor into the estimation of adopting probability in my model. The decay of adopting probability is decided as in the following formula 3:

\[
p_{\text{a}}(t) = p_{\text{a}(t_0)} e^{-\frac{(t-t_0)}{\tau}}
\]

The nodes with high out-degree in social networks represent celebrities. They are active in social networks and more likely to spread new influence to the public. Such nodes have a high spreading probability. However, the probability of adoption is relatively low when people receive influence from such nodes. In contrast, the nodes with low out-degree are average people and they are the main component of most social networks. Such nodes will not publish as many posts as celebrities, but their posts have a bigger influence on their connections. Social Trend also has an influence on the propagation probability. When a new technology was first introduced to the public, it is difficult for people to accept it immediately. As more and more people accept the new technology, the propagation probability will also increase with time. The effect of Social Trend can be estimated by the percentage of active nodes in a social network. The effect of Social Trend will also be added into the estimation of propagation probability.

3.4 Data Sets

In this experiment, two real world data sets from www.arXiv.org, High Energy Physics (hep) data set and Physics (phy) data set, were chosen to test algorithms on the CDE-IC model. The hep data set has 15k nodes and 59k edges, and the phy data set has 37k nodes and 231k edges.

4. EXPERIMENTS

In this paper, I studied how to maximize speed of influence propagation in social networks. I compared four heuristic methods, Random, Distance-Centered, Degree-Centered and my novel Zero-Discount method on the Continuous Dynamic Extended IC (CDE-IC) Model.

I studied the CDE-IC model in two steps. Firstly, I extended the EIC model to Continuous Extended IC (CE-IC) model that studies the continuous influence of an active node. Secondly, I improved the CE-IC model to CDE-IC model, in which the probability of propagation will change with time.
The process of influence propagation is assumed to be discrete, which means each active node can only propagate influence one step further in a unit time. By recording how many nodes are affected in each step, we know both the instant speed and the average speed of the whole process. Each method is run 100 times to obtain an average result.

4.1 Results in CE-IC Model

4.1.1 Comparison of Efficiency

The efficiencies of four algorithms are analyzed by comparing the running time in seed-node selection phase. Seed nodes size is 20 and the time unit is millisecond.

Table 1. Running time of Seed-node Selection Algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>hep</th>
<th>phy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Distance-Centered</td>
<td>13613</td>
<td>186193</td>
</tr>
<tr>
<td>Degree-Centered</td>
<td>45</td>
<td>70</td>
</tr>
<tr>
<td>Zero-Discount</td>
<td>143</td>
<td>230</td>
</tr>
</tbody>
</table>

Random method is the fastest among the four methods. It is only related to size of seed nodes. Distance-Centered is the slowest in these four heuristic methods and does not scale well. It takes O(n) for a root to reach all other nodes. The total time to calculate average distance of all nodes is O(mn). Degree-Centered algorithm is much more efficient than the Distance-Centered algorithm. Time complexity of is O(m + n). My novel Zero-Discount heuristic algorithm is an improvement over the Degree-Centered algorithm. It runs slower than Degree-Centered algorithm to check more nodes, but it is still much efficient than Distance-Centered algorithm. In Zero-Discount algorithm, if a node is selected as a seed node, the degree of all its out-neighbor nodes will be set to zero and put to the end of queue at each step. The time complexity of Zero-Discount is O(mlgk+n).

The experiments in seed-node selection phase show that my novel Zero-Discount algorithm has a comparable efficiency to the Degree-Centered algorithm, and significantly better efficiency than Distance-Centered algorithm. Zero-Discount is about 100 times faster than Distance-Centered method.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>O(k)</td>
</tr>
<tr>
<td>Distance-Centered</td>
<td>O(mn)</td>
</tr>
<tr>
<td>Degree-Centered</td>
<td>O(m+n)</td>
</tr>
<tr>
<td>Zero-Discount</td>
<td>O(mlgk+n)</td>
</tr>
</tbody>
</table>

Random selection is still the slowest. There are two important features in this graph. Firstly, both Degree-Centered and Zero-Discount algorithms have significantly fast instant speed at the initial steps of influence propagation. This is generated by inclusion of high degree nodes in the seed nodes and is the major contributor to the fast average speeds seen in Figure 5. At the later steps of influence propagation, the instant speed of Degree-Centered algorithm declined and was similar to Distance-Centered algorithm.

Secondly, although the instant speeds of Zero-Discount algorithm also declined at later steps, there were several minor spikes during influence propagation such like data point 20 and 30. The difference in instant propagation speed between Degree-Centered and Zero-Discount algorithm is that Degree-Centered algorithm only selects high degree nodes, and there is a large overlapping between the high degree nodes. As the influence propagation proceeds, these high degree nodes add little additional value to the propagation process. In contrast, Zero-Discount algorithm tries to diversify seed nodes selection by eliminating neighbor nodes of seed nodes, and adding high degree nodes from isolated sub-networks as seed nodes. As influence
propagation proceeds, these small isolated sub-networks can be connected. That is why Zero-Discount algorithm performs better than Degree-Centered algorithm at later steps of influence propagation.

![Figure 6: Instant Speed in CE-IC model (hep k=20)](image)

Next, I explored how seed-node size affects the speed of influence propagation. Seed-node size of 50 was used for the same data set hep. The average propagation speeds of the four algorithms are shown in Figure 7. Obviously, increase of seed node size significantly increased propagation speed of all algorithms (Figure 7 vs. Figure 5). Similar to results obtained with a seed-node size of 20, Zero-Discount algorithm has the fastest average propagation speed among the four algorithms. In addition, it is about 20% faster than Degree-Centered algorithm, even better than the result with the seed node size of 20. This result indicates that the drawback of Degree-Centered algorithm is more prominent with increasing of seed-node set. The Zero-Discount algorithm has more advantage over Degree-Centered algorithm in larger seed-node selection.

![Figure 7: Average Speed in CE-IC Model (hep k=50)](image)

The instant propagation speed is shown in Figure 8. It is more obvious that even Zero-Discount loses more speed at the first step, however, Zero-Discount can almost beats Degree-Centered in the following steps.

To test if the faster propagation speed of Zero-Discount algorithm is also true for other data sets, similar experiments were run in a larger data set, the phy data set. The average and instant propagation speeds are shown in Figure 9 and Figure 10, respectively. Consistently, Zero-Discount algorithm has the fastest average and instant propagation speeds among all four algorithms. Distance-centered algorithm performed much worse in this experiment. One reason of this low performance could be that the selected seed nodes have a lower propagation probability.

![Figure 8: Instant Speed in CE-IC Model (hep k=50)](image)

![Figure 9: Average Speed in CE-IC Model (phy k=20)](image)

![Figure 10: Instant speed In CE-IC model (phy k=20)](image)

### 4.2 Results in CDE-IC Model

In the above experiments, Zero-Discount algorithm always has the fastest propagation speed among the four heuristic algorithms in the CE-IC model. Next step, I will add dynamic property into model to test if Zero-Discount is still the best in CDE-IC model when propagation probability changes with time. According to formula 3, the adoption probability of a node will decrease with time in exponential number. I set the half time period to 20, which is half of experiment units. I will also record at which step each active
node tries to propagate influence at initial time for relative edges. I will test CDE-IC model with 50 seed nodes set.

In Figure 11, we can infer similar conclusion as in CE-IC model. We can also see clearly the effect of exponential decreasing in adoption probability. First, the total number of affected nodes in CDE-IC model is about 1/3 less than CE-IC model. Second, the curves in CE-IC model is more like straight line, which means the newly affected nodes is increasing at steady pace. Contrastingly, we can see the decrease in trend of the curve in CDE-IC model. Both models show that my novel Zero-Discount can always beat all the other heuristic methods with best quality of seed nodes set. Finally, we can add a small correction factor to probability if considering Social Trend. However, this value is not significant until a large part of the social networks is under influence. Since we are studying speed of influence propagation, which is more meaningful in a short period, when not so many nodes under influence, we can safely ignore it.

5. CONCLUSION

In this paper, I studied how to maximize speed of influence propagation in social networks. I proposed a new Continuous Dynamic Extended IC (CDE-IC) Model, which is an improved modification of the Extensional Independent Cascade (EIC) Model. The original EIC model has two drawbacks: first, an active node can only try to propagate its influence to its neighbors once; second, the propagation probability does not change between nodes. Both of these problems are solved in my CDE-IC model. I ran four algorithms in CDE-IC Model, Random, Distance-Centered, Degree-Centered and my novel Zero-Discount method. Experiments on two data sets with different sizes of seed nodes all showed that my Zero-Discount method performed better than any other heuristic methods. The previous best method, Degree-Centered, was 10%-20% slower than my method.

6. REFERENCE


