# Analysis of Influence Maximization in Large-Scale Social Networks

Jie Hu*, Kun Meng*, Xiaomin Chen†, Chuang Lin*, Jiwei Huang*
*Department of Computer Science and Technology, Tsinghua University, Beijing, China
†School of Information Engineering, University of Science and Technology Beijing, Beijing, China
jiehu1990@gmail.com, mengkun1024@163.com, chenxiaomin3@gmail.com, chlin@tsinghua.edu.cn, huangjw05@gmail.com

## ABSTRACT

Influence maximization is an important problem in online social networks. With the scale of social networks increasing, the requirements of solutions for influence maximization are becoming more and more strict. In this paper, we discuss two basic methods to compute the influence in general social networks, and then reveal that the computation of influence in series-parallel graph is in linear time complexity. Finally, we propose an novel method to solve influence maximization and show that it has a good performance.

## Keywords

Social Network, Influence Maximization, Series-Parallel Graph

## 1. INTRODUCTION

In recent years, online social networks have received tremendous attention due to its ability in fast, widespread information dissemination. With the popularity of internet, especially the rise of mobile internet and smart phones, social networking has become a finger tap away. As of December 2012, about 70% of online adults use social networking sites in American. And they share texts, pictures and videos in these sites nearly every day. Online social networks are becoming a large information dissemination platform and a huge data warehouse which will have many potential applications.

Viral marketing which is one of the applications in online social networks has become an important topic in both academic and industry areas. Firstly, it selects a set of influential individuals and tries to convince them to adopt the new product or innovation by giving them trial samples or payment. Then, let they recommend the product or innovation to their friends in social networks. Thus, a cascade effect can be triggered by these friends recommending it to their other friends. At last, the new product or innovation can be recommended to a large population of people. However, viral marketing faces several challenges. One of the challenges is selecting most influential user set with the cardinality less than a given number due to the budget, which is also called influence maximization problem. But the evaluation of the influence of a user or a user set is an open problem which has been discussed in many literatures by different methods, such as PageRank based methods [1], user behaviour analysis [12], and classic topology-based heuristics. In this literature, we define the influence by information diffusion models which can explicitly represent the dynamical process of step-by-step information diffusion[8]. *Independent cascade*(IC) model and *linear threshold* model are two basic information diffusion models. We focus on the IC model in this paper. Researchers have proposed many methods to solve the influence maximization problem in IC model. However, these methods can not satisfy the efficiency and accuracy simultaneously. We propose a series-parallel graph based approach to try to solve this problem efficiently and accurately.

The rest of this paper is organized as follows. Section 2 presents the formalization of influence maximization in IC model and related work about solving this problem. Section 3 presents the solution of influence evaluation in the class of series-parallel graph. Section 4 presents our solution of solving influence maximization in general social networks. Section 5 show the performance of our solutions through a simple example.



## 2. PROBLEM DEFINITION AND RELATED WORK

### 2.1 Problem definition and analysis

The IC model is about the step-by-step dynamical process of a message diffusion. In the IC model, we can formalize a given social network as an uncertain directed graph which can be denoted by $\mathcal{G} = (V, E, p)$. The set of vertices $V$ represents the set of users or nodes in the social network. The set of edges $E$ represents the relationship between corresponding users. The function $p : E \to (0 : 1]$, assigns each edge $e = (u, v)$ a probability. It denotes the likelihood of $e$'s existence which represents the probability that $v$ is influenced by $u$. The existence of edges is independent of each other. We call an individual node active if it has adopted the message, and inactive otherwise. We suppose that every individual node has two states, and could only change from inactive to active. It means a node will keep active once activated. Moreover, in IC model, the message disseminates in discrete steps. Let $S_t$ denote the set of nodes that become active at step $t$. Specially, $S_0$ denotes the set of initial active nodes. Then, in IC model, the diffusion of message will proceed in the following process [8][3][2]. The process starts at $t = 1$. At step $t$, every individual node $u \in S_{t-1}$ will activate its every inactive out-neighbor $v$ by the probability of $p(u, v)$. Then it move to the next step. The process will proceed un-

til $S_{t-1}$ become $\emptyset$ where $t$ denotes the current step. In every step, when more than one active nodes attempt to activate the same inactive node, the sequence is arbitrary. And it can be proved the different sequences have the same result.

Then, the influence of a set of nodes in IC model, denoted by $\sigma(A)$, can be defined as follows.

DEFINITION 1. *[8] The influence of a set of nodes A is defined to be the expected number of active nodes at the end of the information dissemination process which selects A as the initial active set.*

Kempe [8] first prove that the influence maximization problem in independent cascade model is NP-hard and then propose a greedy algorithm (algorithm 1) which has approximation guarantee.

**Algorithm 1** Greedy($k$)

1: $A \leftarrow \emptyset$
2: **while** $|A| < k$ **do**
3: $\quad a \leftarrow arg\ max_{a \in V}\{\sigma(A \bigcup \{a\}) - \sigma(A)\}$
4: $\quad A \leftarrow A \bigcup \{a\}$
5: **end while**

It is proved that the function $\sigma(A)$ is non-negative, monotone, and submodular. So we can conclude that the approximate factor of the simple greedy algorithm is $(1-1/e)$ if the $\sigma(A)$ is calculated exactly [13]. However, given a user set $A$, the calculation of $\sigma(A)$ in IC model is NP-hard. And Kempe use Monte-Carlo simulation to estimate the $\sigma(A)$. Thus, the approximate factor of the total algorithm is $(1 - 1/e - \epsilon)$, where $\epsilon$ depends on accuracy of the simulation.

According to the definition of influence, $\sigma(A)$ is identical to the expected number of nodes which the nodes of $A$ can reach in the corresponding uncertain graph. Similar to the computation of reliability or reachability [7], $\sigma(A)$ can be computed in two methods.

The first is the possible graph based method. For a given uncertain graph $\mathcal{G} = (V, E, p)$, there are a total of $2^{|E|}$ possible graphs. Let $G = (V_G, E_G)$ denote a possible graph of $\mathcal{G} = (V, E, p)$. Then the existence probability $Pr[G]$ of the possible graph $G$ is as follows.

$$Pr[G] = \prod_{e \in E_G} p(e) \times \prod_{e \in E \setminus E_G} (1 - p(e)) \tag{1}$$

Let $\sigma_G(A)$ be the number of nodes that $A$ can reach in $G$. Then $\sigma(A)$ can be computed as follows.

$$\sigma(A) = \prod \sigma_G(A) \times Pr[G] \tag{2}$$

The value of $\sigma_G(A)$ can be computed by the DFS algorithm in graph $G$ from $A$.

Although we can compute the $\sigma(A)$ by equation 2, it has the complexity of $\Theta(|A| \times |V| \times 2^{|E|})$.Thus, Monte-Carlo simulation is adopted to approximate the $\sigma(A)$ [8] .

The other method is the path based method. We need to find out all the simple paths from $A$ to other nodes in uncertain graph $\mathcal{G}$. Let $P_1^v, P_2^v, \cdots, P_{M(v)}^v$ denote all such paths that end with node $v$, where $M(v)$ denotes the number of all such paths end with $v$. Let $\sigma^v(A)$ denote the influence of $A$ on $v$ which is equal to the probability that $A$ can reach $v$ in uncertain graph $\mathcal{G} = (V, E, p)$. Then we can compute $\sigma(A)$ as follows.

$$\sigma(A) = \sum_{v \in V} \sigma^v(A) \tag{3}$$

Let the stochastic variable $X_i^v$ be:

$$X_i^v = \begin{cases} 1, & \text{the path } P_i^v \text{ exists,} \\ 0, & \text{others .} \end{cases} \tag{4}$$

Thus,

$$Pr\{X_{i_1}^v \wedge X_{i_2}^v \wedge \ldots \wedge X_{i_n}^v\} = \prod_{e \in P_{i_1}^v \bigcup \cdots \bigcup P_{i_n}^v} p(e) \tag{5}$$

The $\sigma^v(A)$ can be computed as follows.
(a) For $v \notin A$

$$\begin{aligned} \sigma^v(A) &= Pr\{X_1^v \vee X_2^v \vee \ldots \vee X_{M(v)}^v = 1\} \\ &= \sum_{i=1}^{M(v)} Pr\{X_i^v = 1\} \\ &\quad - \sum_{i \neq j} Pr\{X_i^v \wedge X_j^v = 1\} \\ &\quad + \ldots + (-1)^{M(v)-1} Pr\{\bigwedge_{i=1}^{M(v)} X_i^v = 1\} \end{aligned} \tag{6}$$

(b)For $v \in A$, $\sigma^v(A) = 1$.

Let $f(x) = \sum_{i=1}^{x}(\prod_{j=1}^{i}(x - i))$. Then we can get that the total number of simple paths from $A$ to the other nodes is $O(f(|V| - |A|))$. So the complexity of the path based method is $O((|V| - |A| \times 2^{|A| \times f(|V|-|A|)}))$, which implies that the complexity maybe even higher than possible graph based methods. However, the path based methods provide important ideas to design heuristic algorithms [2] [9].

## 2.2 Related Work

Domingos et al. [4] first propose the influence maximization problem. They model the interaction of users as a markov random field and provide heuristics to choose users who have large influence in network. Kempe et al.[8] formulate the problem as a discrete optimization problem and propose a approximate algorithm(algorithm 1). However, the approximate algorithm is time-consuming. Recently, researchers have improved this algorithm mainly in two ways. One is reducing the number of $\sigma(A)$ calls. The other is improving the efficiency of calculating $\sigma(A)$. Leskovec et al.[11] propose an improvement approach which is called CELF. The idea is as follows. Let $A_i$ denote the set $A$ in iteration $i$ in algorithm 1. Let $A_j$ denote the set $A$ in iteration $j$. Suppose that $i < j$. Then, we can get $\sigma(A_j \bigcup \{a\}) - \sigma(A_j) \geq \sigma(A_i \bigcup \{a\}) - \sigma(A_i)$. That means, if a node $x$ has satisfied $\sigma(A_i \bigcup \{x\}) - \sigma(A_i) > \sigma(A_j \bigcup \{a\}) - \sigma(A_j)$, we do not need to compute $\sigma(A_i \bigcup \{a\})$ in iteration $i$. Thus, the number of $\sigma(A)$ calls can be reduced significantly. Goyal et al.[5] propose an extension to CELF, called CELF++, which can reduce the number of $\sigma(A)$ calls further. The idea is that in any iteration, when the $\sigma(A \bigcup \{a\})$ and $\sigma(A \bigcup \{x\} \bigcup \{a\})$ can be computed simultaneously without much extra overhead, where $\sigma(A \bigcup \{x\})$ is maximal until now in the current iteration,they should be computed at the same iteration. Thus,

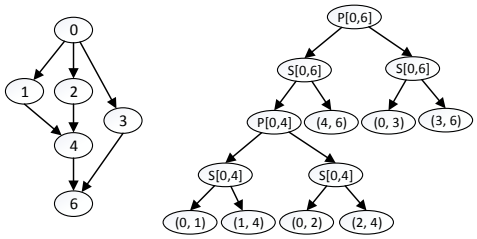


**Figure 1: A series-parallel graph (left) and its decomposition tree (right)**

**Algorithm 2** $\sigma\_SPG(T)$

1: **if** $T$ has one node **then**
2: Return $p(T)$
3: **end if**
4: Let $LC$ be the left child of $T$
Let $RC$ be the right child of $T$
5: **if** the root of $T$ is a $S$ node **then**
6: Return $\sigma\_SPG(LC) \times \sigma\_SPG(RC)$
7: **else**
8: Return $1 - (1 - \sigma\_SPG(LC)) \times (1 - \sigma\_SPG(RC))$
9: **end if**

if the current iteration chooses $x$ finally, we do not need to compute $\sigma(A \bigcup \{a\})$ in the next iteration.

Kimura et al. [10] utilize the strongly connected component (SCC) to improve the simple algorithm method. Specially, they use these two properties: (1) if $u$ and $v$ are in the same SCC in possible graph $G$, then $\sigma_G(A \bigcup \{u\}) = \sigma_G(A \bigcup \{v\})$; (2) if $\exists a \in A$ and $a$ is in the same SCC with $u$, then $\sigma_G(A \bigcup \{u\}) = \sigma_G(A)$.

Although many researchers have proposed improvements for the simple greedy algorithm, they are not efficient enough for the large scale of current social networks. Thus, the heuristic algorithms are proposed. Kimura et al. [9] propose two methods based on shortest path which are called SPM and SP1M. Chen et al. propose degree discount heuristics [3] and maximum influence path based heuristics [2] which is called PMIA. However, the accuracy of these methods needs to be improved.

# 3. INFLUENCE COMPUTATION IN SERIES-PARALLEL GRAPH

We will discuss about the influence computation in series-parallel graph in this section.

The series-parallel graph $G$ [6] is a special class of directed graphs. It has one source and one destination. The definition is as follows.

(1) The graph which has two vertexes and a single directed edge connecting them is the simplest series-parallel graph.

(2) If $G_1(V_1, E_1)$ is a series-parallel graph from $s_1$ to $t_1$, and $G_2(V_2, E_2)$ is a series-parallel graph from $s_2$ to $t_2$. Then, if $G_1$ and $G_2$ satisfy either of the following two conditions, the graph $G_1 \bigcup G_2$ is a series-parallel graph, where $G_1 \bigcup G_2$ denotes the graph $G(V_1 \bigcup V_2, E_1 \bigcup E_2)$.

(a)$G_1$ and $G_2$ are connected in series way which requires $V_1 \bigcap V_2 = \{t_1\} = \{s_2\}$ and $E_1 \bigcap E_2 = \emptyset$.

(b)$G_1$ and $G_2$ are connected in parallel way which requires $V_1 \bigcap V_2 = \{s_1, t_1\}$, $s_1 = s_2$, $t_1 = t_2$, and $E_1 \bigcap E_2 = \emptyset$.

The series-parallel graph can be represented by the decomposition tree conveniently[6]. A simple example about the series-parallel graph and its decomposition tree is as figure 1 shown. The decomposition tree is a binary tree. Each node represents a series-parallel subgraph. The leaf nodes represent the edges of the series-parallel graph. And the inner nodes represent the subgraphs which are composed by two of their child nodes. Each inner node has a type which represent the composition type which can be a series composition or a parallel composition. For notational simplicity, we use $S$ to denote the series node and $P$ to denote the parallel node.

Let a social network be $\mathcal{G} = (V, E, p)$ which has the series-parallel topology. For any given $A(A \subseteq V)$, we can calculate $\sigma^v(A)(v \in V \setminus A)$ as follows.

(a) If $|A| = 1$ (let $A = \{a\}$), construct a maximum series-parallel subgraph $\mathcal{G}_{a,v}$ with source $a$ and destination $v$ from $\mathcal{G}$ which can be easily realized by the decomposition tree. Let $T_{a,v}$ be the decomposition tree of $\mathcal{G}_{a,v}$. Then we can use algorithm 2 to compute $\sigma^v(A)$.

(b) If $|A| > 1$, in other words, there is more than one nodes in $A$, we can convert this case to (a) by theorem 1.

THEOREM 1. *Let a social network be $\mathcal{G} = (V, E, p)$. For a set $A$ with $|A| > 1$ and $A \subseteq V$, we can create a new uncertain graph $\mathcal{G}^{\#}$: add a new node $u_0$ to $\mathcal{G}$ and connect $u_0$ with every node $v$ of $A$ by adding edge $(u_0, v)$ with probability 1 to $\mathcal{G}$. Then the influence of $A$ in $\mathcal{G}$ is equal to the influence of $u_0$ on $V$ in $\mathcal{G}^{\#}$.*

The complexity of algorithm 2 is a linear function of the size of $T$, so the computation of $\sigma(A)$ is efficient in series-parallel graph.

# 4. INFLUENCE MAXIMIZATION BASED ON SERIES-PARALLEL SUBGRAPH

In the previous section, we show that computing influence in a series-parallel graph is easy. However, real social networks are not series-parallel graphs typically. In order to get an efficient influence maximization approach in general social networks, we need to extract a maximum influence series-parallel subgraph (MSPG) for every two nodes $u$ and $v$. However, as far as we know, the construction of the maximum influence series-parallel subgraph from $u$ to $v$ is an open problem. Thus, we give a heuristic algorithm FindMSPG to solve this problem. The computation of the set of connectible vertexes and the definition of augmenting path in algorithm 3 can be referred to [6]. However, using FindMSPG to extract MSPG is not efficient. It is because that we have to call the FindMSPG function for every pair $(u, v)$. In fact, we can extract MSPGs form $u$ to any other nodes simultaneously which can improve the efficiency greatly. As space is limited, we will not discuss about it here.

When get the series-parallel sub-graphs, we can compute influence efficiently. Then, the influence maximization problem can be solved.

# 5. EXPERIMENTS: A SIMPLE EXAMPLE

We will compare our solution with these solutions (PMIA, SPM, SP1M) which can compute the influence nearly in linear time. Although this example is simple, it can reveal the advantages of our solution. As figure 2 shown, considering a

**Algorithm 3** FindMSPG($\mathcal{G}, u, v$)

1: Let $P$ be the maximum influence path from $u$ to $v$
2: $G \leftarrow P$
3: Let $U$ be the set of connectible vertex pairs in $G$
4: **while** $U \neq \emptyset$ **do**
5: Selecting vertex pair $(x, y)$ which has the maximum probability to improve the influence from $u$ to $v$.
6: Let $P$ be the maximum influence valid augmenting path between $u$ and $v$ with respect to $G$ and $\mathcal{G}$
7: Add $P$ to the $G$
8: Recompute $U$
9: **end while**

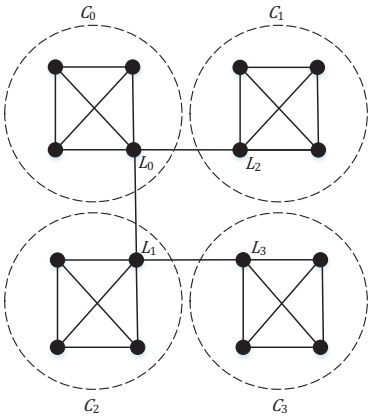


**Figure 2: A simple example of social network**

simple example, a social network is composed of four communities (denoted by $C_0, C_1, C_2, C_3$ respectively), and each community has 100 members (In figure 2, we draw only four members for each community). In a community, each two member are friends. In other words, each community is fully connected. Every community has one leader (denoted by $L_0, L_1, L_2, L_3$ respectively) and the communication between communities can only through the leaders. The connections of the four communities are as figure 2 shown. Let the activation probability be 0.2. Then, the results of $\sigma(L_0)$ are as table 1 shown.

It can be proved that both the results of MSPG and MIA are smaller than the real value. So, the bigger one is closer to the real value. Besides, we can get an upper bound of the real value is 144. Therefore, we can get the real value is between 142 and 144. Thus, our result is closer to the real value and the accuracy rate is more than 98.6%.

# 6. FUTURE WORKS

**Table 1: The Influence of $L_0$ in four communities and total network**

| | $C0$ | $C1$ | $C2$ | $C3$ | Total |
|---|---|---|---|---|---|
| MSPG | 98.55 | 19.71 | 19.71 | 3.94 | 142 |
| PMIA | 20.80 | 4.16 | 4.16 | 0.83 | 30 |
| SPM | 20.80 | 4.16 | 4.16 | 0.83 | 30 |
| SP1M | 98.55 | 53.94 | 53.94 | 15.31 | 222 |
| Bound | 100 | 20 | 20 | 4 | 144 |

We will use real social-network data to test the performance of our solution, and then improve it. Two possible directions of our future work are as follows.

(1) Construct an information diffusion model which is closer to the real-life, such as taking information content and time factor into account.

(2) Realize our solution in the cloud platforms to make further efforts to improve the performance.

# 7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China(61071065).

# 8. REFERENCES

[1] A. Balmin, V. Hristidis, and Y. Papakonstantinou. Objectrank: authority-based keyword search in databases. VLDB'04, pages 564–575.

[2] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD'10, pages 1029–1038, New York, USA.

[3] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. KDD'09, pages 199–208, New York, USA.

[4] P. Domingos and M. Richardson. Mining the network value of customers. KDD'01, pages 57–66, New York, USA.

[5] A. Goyal, W. Lu, and L. V. Lakshmanan. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proceedings of the 20th international conference companion on World wide web*, pages 47–48, New York, USA, 2011.

[6] P. Hintsanen and H. Toivonen. Finding reliable subgraphs from large probabilistic graphs. *Data Mining and Knowledge Discovery*, 17:3–23, 2008.

[7] R. Jin, L. Liu, B. Ding, and H. Wang. Distance-constraint reachability computation in uncertain graphs. *Proc. VLDB Endow.*, 4(9):551–562, June 2011.

[8] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. KDD'03, pages 137–146, New York, USA.

[9] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In *Knowledge Discovery in Databases: PKDD 2006*, volume 4213, pages 259–271.

[10] M. Kimura, K. Saito, and R. Nakano. Extracting influential nodes for information diffusion on a social network. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1371, 2007.

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. KDD'07, pages 420–429, New York, USA.

[12] Y. Li, C. Lai, and C. Chen. Discovering influencers for marketing in the blogosphere. *Information Sciences*, 181(23):5143 – 5157, 2011.

[13] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.