

# Influence Maximization for Cascade Model with Diffusion Decay in Social Networks

Zhijian Zhang<sup>1,2</sup>, Hong Wu<sup>1,3</sup>, Kun Yue<sup>1(✉)</sup>, Jin Li<sup>4</sup>, and Weiyi Liu<sup>1</sup>

<sup>1</sup> School of Information Science and Engineering, Yunnan University, Kunming, China  
kyue@ynu.edu.cn

<sup>2</sup> College of Science, Kunming University of Science and Technology, Kunming, China

<sup>3</sup> College of Computer Science and Engineering, Qujing Normal University, Qujing, China

<sup>4</sup> School of Software, Yunnan University, Kunming, China

**Abstract.** Maximizing the spread of influence is to select a set of seeds with specified size to maximize the spread of influence under a certain diffusion model in a social network. In the actual spread process, the activated probability of node increases with its newly increasing activated neighbors, which also decreases with time. In this paper, we focus on the problem that selects  $k$  seeds based on the cascade model with diffusion decay to maximize the spread of influence in social networks. First, we extend the independent cascade model to incorporate the diffusion decay factor, called as the cascade model with diffusion decay and abbreviated as CMDD. Then, we discuss the objective function of maximizing the spread of influence under the CMDD, which is NP-hard. We further prove the monotonicity and submodularity of this objective function. Finally, we use the greedy algorithm to approximate the optimal result with the ration of  $1 - 1/e$ .

**Keywords:** Social networks · Influence maximization · Cascade model · Diffusion decay · Submodularity · Greedy algorithm

## 1 Introduction

With the popularity of online social networks, such as Facebook, Twitter, WeChat, etc., the online social networks play an increasingly important role in daily communication among people. Many researchers have studied the diffusion phenomenon in social networks, such as the diffusion of news and opinions [1, 2], the adoption of products [3], the spread of infectious diseases [4–6], etc. Influence maximization is a fundamental problem of the diffusion in social networks. An application of influence maximization is viral marketing [3, 7, 8]. There have been extensive commercial instances of viral marketing succeed in real life, such as Nike Inc., used [orkut.com](http://orkut.com), and [facebook.com](http://facebook.com) to market products successfully [9] and the Hotmail phenomenon [10].

Focusing on how to model the diffusion process, some researchers have proposed various diffusion models for the diffusion of innovations, ideas, etc. [12–15]. Randomness, the cumulative effect and the decay characteristic are the main characteristics of propagation. Most of the existing models describe the first two characteristics. But few researchers focus on the decay characteristics of influence diffusion. In short, the

diffusion decay refers to the decay of influence during the diffusion. For example, Tom read an interesting piece of news, and then he may forward it to his friends with probability  $p$  on the first day. But if not, he still may forward it with probability  $p'$  on the second day ( $p' < p$ ), and  $p'$  decreases with time. That is, in the real diffusion process, the influence will decrease with time going and is reflected by the decreasing of the activate probability of node. Thus, a model that does not consider diffusion decay cannot simulate the actual spread process well. Furthermore, it is critical to model the spread process with diffusion decay for analyzing the influence maximization problem of social networks, which exactly we will solve in this paper.

In our study, we focus on the problem of selecting the seeds to maximize the influence spread considering diffusion decay in a social network. For this purpose, we consider the following problems:

- (1) How to model the spread process with diffusion decay?
- (2) How to select  $k$  seeds to maximize the influence spread?

For the problem (1), it is natural to consider extending the classic independent cascade model (IC model) [8] to incorporate the influence probability decaying with time, which is called as cascade model with diffusion decay, abbreviated as CMDD. In the CMDD, the activate probability is influenced by the following three factors: the previous cumulative effects, the influence power of new activated neighbor nodes and the decay factor. The probability of node  $v$  is a function of these three factors, which can well reflects realistic characteristics of influence spread in a social network.

For the problem (2), selecting  $k$  seeds to maximize the influence spread under the CMDD is NP-hard. Whether the CMDD model defined upon the IC model still keeps the monotonicity and submodularity is the key and difficult part in our work. We prove the monotonicity and submodularity of the objective function, and thus the greedy algorithm can be used to approximate the optimal result with the ration of  $(1 - 1/e)$  based on the theoretic conclusion given by Nemhauser et al. [11].

In order to test the feasibility of the method proposed in this paper, we implement our algorithms and make corresponding experiments.

The reminder of this paper is organized as follows. In Sect. 2, we introduce related work. In Sect. 3, we give CMDD to model the influence spread of node. In Sect. 4, we obtain the objective function of influence maximization under the CMDD, and prove the monotonicity and submodularity of this objective function. In Sect. 4.2, we exploit the approximation algorithm to maximize the influence spread. In Sect. 5, we show the experimental results and performance studies. Finally in Sect. 6, we conclude and discuss the further work.

## 2 Related Work

Domingos et al. [7] discussed the influence maximization as an algorithm problem for the first time, and they modelled custom network as a graph and used a Markov random field to calculate the influence probabilistic among them. In the aspect of modeling the

diffusion process of influence, many researchers proposed various methods of influence maximization from various perspectives [8, 12–15].

Kempe et al. [8] formulated the problem of selecting a set of influence individuals to maximize the influence spread as a discrete optimization problem and proposed independent cascade model (IC model) and linear threshold model (LT model) based on earlier works [16–19]. The key feature of the model is that diffusion events along every arc in the social graph are mutually independent [20]. The LT model reflects the influence cumulative effect during the process of propagation, and the IC model can reflect the randomness of node activation. In this paper, our CMDD is based on the IC model, and it not only retains the cumulative effect in LT model, but also describes the influence decay during the diffusion. Our CMDD retains the monotonicity and submodularity in both LT and IC models.

Saito et al. [12] presented a method for predicting diffusion probabilities by using the Expectation Maximization algorithm based on the IC model. Yang and Leskovec [14] presented the linear influence model (LIM) to model the global influence of nodes. Goyal et al. [13] proposed three models: static model, continuous time model, discrete time model, in which the influence probabilities are relative to the action log instead of the discrete time step. In their works, dynamic activate probability is not discussed. In our work, we employ *Inf* to describe node influence power that can reflect not only the graphic characteristics but also some actual factors. The node's activate probability is changing with recently activated neighbours and the decay factor during the diffusion.

In the time-critical influence maximization problem, Chen et al. [23] extended the IC model and the LT model to incorporate the time delay aspect of influence diffusion, but the diffusion decay is not considered. Liu et al. [24] defined time constrained activate probability which is an assumed value at different times. In CMDD, we mainly consider that the influence diffusion probabilities of nodes decay with varying time step. Actually, the probability at time  $t$  is a function of the probability at time  $t-1$ , the decay factor and the influence by neighbours which are activated at  $t-1$ , which is dynamic.

In the aspect of how to select seeds, many researches proposed heuristics and tried to solve the influence problem more efficiently [8, 21, 22]. In terms of algorithm design, our work follows the idea given in [8]. To select the optimal seeds is NP-hard under the CMDD, and then the greedy algorithm is used to approximate the optimal result based on the mathematical theory given in [11].

### 3 Cascade Model with Diffusion Decay

A social network is denoted as an undirected graph  $G = (V, E)$ , where  $V$  is the set of nodes representing individuals and  $E$  is the set of edges representing the relationships among individuals. There are two classic diffusion models. One model describing how influence spreads in social network is LT model [8], which considers the influence accumulation of diffusion with time steps. Another model is IC model, in which an activated node  $u$  tries to activate its neighbor  $v$  with initialized  $p_{uv}$  only once [8].

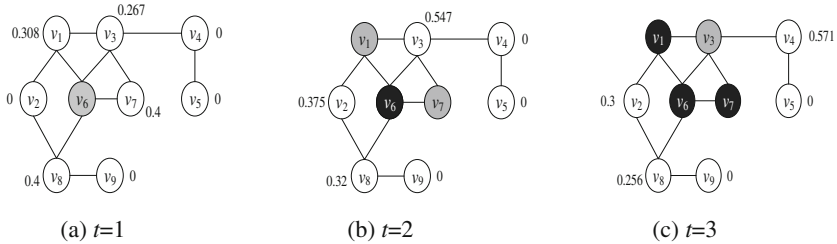
In this paper, we propose the CMDD based on the IC model. CMDD combines the time step characteristics of influence diffusion and the influence accumulation. In this

model, each node is either active or inactive. At step  $t$ , the node  $v$  is activated with probability  $p'_v$ , which can be described as follows:

$$p'_v = \alpha \times p_v^{t-1} + \frac{\sum_{w \in A_{t-1} \cap N(v)} Inf_w}{Inf_v + \sum_{u \in N(v)} Inf_u} \tag{1}$$

where  $A_{t-1}$ ,  $N(v)$  and  $\alpha$  denote the activated nodes at step  $t-1$ , the neighbors of node  $v$  and the decay parameter of influence respectively, where  $0 \leq \alpha \leq 1$ . This decay parameter can be denoted as a constant or an exponential function with parameters depending on the time. In order to facilitate the discussion, we employ a constant to denote the decay parameter. For  $\alpha = 0$ , this model is similar to the IC model. For  $\alpha > 0$ , this model can reflect the random property and the influence accumulation of the LT model. The greater the value of  $\alpha$ , the slower the process of influence decay.  $Inf_w$  denotes the influence power of node  $w$ , such as the node's importance degree.  $N(v)$  denotes neighbors of  $v$  and  $A_{t-1}$  denotes the nodes set which are activated at time  $t-1$ .

**Example 1.** Figure 1 shows an example of the diffusion process of CMDD. We assume  $\alpha = 0.8$  and  $Inf_v$  is the degree of node  $v$ . Initially at  $t = 0$ , one seed  $v_6$  is activated. At step  $t = 1$ ,  $v_6$  tries to activate its inactive neighbors with probabilities  $p_{v_1}^1 = 0.308$ ,  $p_{v_3}^1 = 0.267$ ,  $p_{v_7}^1 = 0.4$  and  $p_{v_8}^1 = 0.4$  respectively. At step  $t = 2$ ,  $v_1$  and  $v_7$  are randomly activated, but  $v_3$  and  $v_8$  are inactive, and then the activated probabilities of  $v_2$ ,  $v_3$  and  $v_8$  are  $p_{v_2}^2 = 0.375$ ,  $p_{v_3}^2 = 0.547$  and  $p_{v_8}^2 = 0.32$ . Similarly, we can obtain the activated probabilities of nodes at step  $t = 3$ .



**Fig. 1.** The diffusion process of the CMDD ( $\alpha = 0.8$ ). In (a), (b) and (c), grey nodes denote the newly activated nodes, and the black nodes denote the activated nodes that lose the activated ability, other nodes are inactive.

### 4 Maximizing Influence Spread Under CMDD

In this section, we define the objective function of influence maximization problem under the CMDD, which is NP-hard. Then, we show that the objective function is monotone and submodular, which leads to a greedy approximation based on the theory given by Nemhauser et al. [11].

### 4.1 Objective Function of Influence Maximization Problem

The influence maximization problem is an optimal problem, in which given a graph  $G = (V, E)$ , the number of the seed  $k$ , we want to find a seed set  $S$  of the size  $k$  such that the expected number of nodes is maximized. Now, we first consider the objective function of influence maximization problem.

At step  $t = 0$ ,  $A_0(S) = S$ , the expected activated value of influence maximization under the CMDD is  $E_{t=0}(S) = A_0(S)$ . We can obtain the expected activated value at step  $t$  as follows:

$$E_t(S) = \alpha \times E_{t-1}(S) + \sum_{i \in V \setminus A_{t-1}(S)} \frac{\sum_{k \in A_{t-1}(S) \cap N(i)} Inf_k}{Inf_i + \sum_{j \in N(i)} inf_j} \tag{2}$$

The overall expected activated values in  $t$  steps is equal to the sum of the expected activated value with  $t$  steps, that is,

$$E(S) = \sum_{t=0}^t E_t(S) = \sum_{t=0}^t (\alpha \times E_{t-1}(S) + \sum_{i \in V \setminus A_{t-1}(S)} \frac{\sum_{k \in A_{t-1}(S) \cap N(i)} Inf_k}{Inf_i + \sum_{j \in N(i)} inf_j}) \tag{3}$$

To select the optimal seed set to maximize the influence spread with the objective function and under the CMDD is NP-hard. We can prove the monotonicity and submodularity of the objective function.

Obviously, we have

$$E_t(S \cup \{u\}) \geq E_t(S) \tag{4}$$

Thus, the objective function  $E(S)$  is monotone.

We now prove the submodularity of objective function  $E(S)$ .

**Theorem 1.** The objective function is submodular, if for all subsets  $S_1 \subseteq S_2 \subseteq V$  and  $u \in V \setminus S_2$ , we have  $E(S_1 \cup \{u\}) - E(S_1) \geq E(S_2 \cup \{u\}) - E(S_2)$ .

**Proof.** We employ the Mathematical Induction to prove Theorem 1.

At step  $t = 1$ , the objective function  $E(S)$  is obviously submodular.

At step  $t - 1$ , if the objective function is submodular, then we have

$$E_{t-1}(S_1 \cup \{u\}) - E_{t-1}(S_1) \geq E_{t-1}(S_2 \cup \{u\}) - E_{t-1}(S_2) \tag{5}$$

At step  $t$ , we have

$$\begin{aligned}
 & E_t(S_1 \cup \{u\}) - E_t(S_1) \\
 &= \alpha(E_{t-1}(S_1 \cup \{u\}) - E_{t-1}(S_1)) + \sum_{i \in V \setminus A_{t-1}(u)} \frac{\sum_{k \in A_{t-1}(u) \cap N(i)} Inf_k}{Inf_i + \sum_{j \in N(i)} inf_j} - \left( \sum_{i \in V \setminus A_{t-1}(u \cap S_1)} \frac{\sum_{k \in A_{t-1}(u \cap S_1) \cap N(i)} Inf_k}{Inf_i + \sum_{j \in N(i)} inf_j} \right) \quad (6)
 \end{aligned}$$

We have the similar expression of  $E_t(S_2 \cup \{u\}) - E_t(S_2)$ . We can see the activated process as flip the coin. Based on Equality (5) and (6), we have

$$E_t(S_1 \cup \{u\}) - E_t(S_1) \geq E_t(S_2 \cup \{u\}) - E_t(S_2) \quad (7)$$

The linear combination of a submodular function is also submodular, so we have  $E(S_1 \cup \{u\}) - E(S_1) \geq E(S_2 \cup \{u\}) - E(S_2)$ .

### 4.2 Greedy Algorithm for Influence Maximization Problem

We have proven that the objective function of influence maximization problem under the CMDD is monotone and submodular. According to the result proposed in [11], the greedy algorithm given in Algorithm 1 can be used to approximate the optimal result with the relation of  $1 - 1/e$ . The algorithm selects the node that provides the largest marginal gain to the seed set, and each time one node will be selected as a seed.

---

**Algorithm 1:** Greedy ( $k, S$ )

---

**Input:** A social network:  $G=(V, E)$ , number of seeds:  $k$ ,

**Output:** Seed Set  $S$ ,

**Steps:**

- 1: initialize  $S=\Phi$
  - 2: for  $i=1$  to  $k$  do
  - 3:   select  $u = \operatorname{argmax}_{v \in V \setminus S} (E(S \cup \{v\}) - E(S))$
  - 4:    $S=S \cup \{u\}$
  - 5: end for
  - 6: return  $S$
- 

The running time of Algorithm 1 is determined by the greedy part at step 3. The time complexity of Algorithm 1 is  $O(knd_1d_2)$ , where  $k$  is the number of seeds,  $n$  is the number of nodes in network,  $d_1$  is the average degree of nodes and  $d_2$  is the max distance from other inactive nodes to node  $v$ , when we calculate the contribution of  $v$ .

## 5 Experimental Results

To test the feasibility and effectiveness of selecting seeds under the time cascade decay model, we implemented our method and made corresponding performance studies.

### 5.1 Experimental Setup

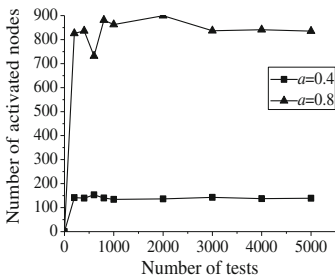
The ca-HepTh and ca-GrQc are HEP-TH (High Energy Physics-Theory) collaboration network extracted from the e-print (<http://arXiv.org/>). The former is extracted from the “High Energy Physics” and the latter is extracted from the “General Relativity”. The nodes in these two networks are *authors* and an edge between two nodes means the *two coauthored at least one paper*. The p2p-Gnutella08 record the Gnutella peer to peer network from August 8 2002, where nodes represent hosts in the Gnutella network topology and edges represent connections between the Gnutella hosts (Table 1).

**Table 1.** Statistics of the two real-world networks in resulting graph

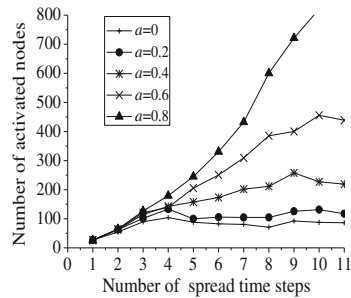
Dataset	ca-HepTh	ca-GrQc	p2p-Gnutella08
Number of nodes	9878	5242	6301
Number of edges	51971	28980	20777

### 5.2 Performance Studies

First, we tested the convergence rate of influence spread in ca-HepTh. In this experiment, we tested the influence spread with  $\alpha = 0.4$  and  $\alpha = 0.8$  under the CMDD respectively where spread time steps  $t = 5$  and node  $user_{ID} = 1441$  with high degree for obvious experiment result. Figure 2 shows that the convergence rate with  $\alpha = 0.4$  was faster than  $\alpha = 0.8$ , and the number of the convergence of influence spread with  $\alpha = 0.4$  and  $\alpha = 0.8$  were 1000 and 3000 respectively. This was because that the influence accumulation of node decreased slowly when the value of  $\alpha$  is greater.



**Fig. 2.** The convergence rate of influence spread with different  $\alpha$  value



**Fig. 3.** The expectation value of active nodes with different  $\alpha$  values

Then, we tested the relationship of the expectation value of influence spread with different  $\alpha$  in ca-HepTh. We compared the expectation value of influence spread with  $\alpha = 0.2$ ,  $\alpha = 0.4$ ,  $\alpha = 0.6$ ,  $\alpha = 0.8$  and  $\alpha = 0$ , where we assigned the spread time step from 1 to 10 and node  $user_{ID} = 63113$ . The comparison is shown in Fig. 3. It can be seen that greater  $\alpha$ , the greater expectation value under the CMDD, since the value of  $\alpha$  is greater, the value of influence probabilities of nodes decreases slower.

It is known that the max-degree algorithm [25] is well regarded as the effective algorithm for the networks with power law distributions, and it sorts the nodes by the degree, and it selects  $k$  max degree nodes as seeds. Random algorithm selects  $k$  seeds randomly. Finally, we tested the effectiveness of Algorithm 1. In this experiment, we selected 20 seeds with Algorithm 1 with Depth = 1 and Depth = 2, the max-degree algorithm (denoted as Max-degree) and random algorithm (denoted as Random) to maximize the influence spread in ca-GrQc and p2p-Gnutella08 and set  $\alpha = 0.4, t = 3$ . The depth in greedy algorithm means the max nodes distance we consider. If Depth = 1, we only consider the neighbors of active nodes. If Depth = 2, we consider not only the neighbors of active nodes but also the neighbors of their neighbors. Figure 4(a) shows that the greedy algorithm (denoted as Greedy) is better than Max-degree and outperforms Random. But in Fig. 4(b) and (c), the greedy algorithm is close to Max-degree, since the  $Inf_v$  of node  $v$  is calculated by degree in our experiments, which verifies that our proposed CMDD model and the corresponding algorithm are feasible.

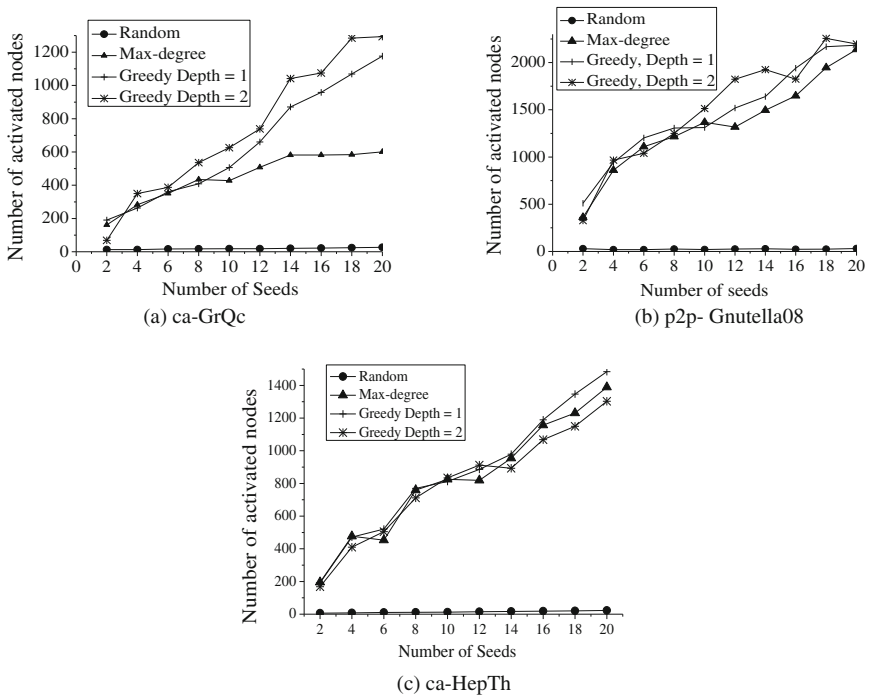


Fig. 4. Expectation value of active nodes with different algorithms inca-GrQc, p2p-Gnutella08 and ca-HepTh,

## 6 Conclusions and Future Works

In this paper, we redefined the node activate probability and proposed the CMDD, which is close to the real diffusion process. The CMDD reflects the change of probability with



time step and new activated nodes, meanwhile it retains the cumulative effect and randomness. Then we proved the monotone and submodularity of this objective function and the greedy algorithm is used to approximate the optimal result.

However, our algorithm is not far superior to max-degree algorithm on some datasets. It is because the  $Inf_v$  of node  $v$  is calculated by degree in our experiments. We will extend our experiments to some real networks in which the  $Inf_v$  is determined by some actual factor. Furthermore, employing a constant to describe the diffusion decay parameter has its limitations. The decay factor function that can better describe the real spread process in a social network is still worth discussing. These are our next research directions.

**Acknowledgement.** This paper was supported by the National Natural Science Foundation of China (61562091), Natural Science Foundation of Yunnan Province (2014FA023, 201501CF00022), Program for Innovative Research Team in Yunnan University (XT412011), and Program for Excellent Young Talents of Yunnan University (XT412003).

## References

1. Gruhl, D., Guha, R., Liben-Nowell, D., et al.: Information diffusion through blogspace. In: WWW, pp. 491–501 (2004)
2. Leskovec, J., Krause, A., Guestrin, C., et al.: Cost-effective outbreak detection in networks. In: KDD, pp. 420–429 (2007)
3. Datta, S., Majumder, A., Shrivastava, N.: Viral marketing for multiple products. In: ICDM, pp. 118–127 (2010)
4. Bailey, N.T.J.: *The Mathematical Theory of Infectious Diseases and Its Applications*. Haffner Press, Royal Oak (1975)
5. Anderson, R.M., May, R.M., Anderson, B.: *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, Oxford (1992)
6. Kim, L., Abramson, M., Drakopoulos, K., Kolitz, S., Ozdaglar, A.: Estimating social network structure and propagation dynamics for an infectious disease. In: Kennedy, W.G., Agarwal, N., Yang, S.J. (eds.) SBP 2014. LNCS, vol. 8393, pp. 85–93. Springer, Heidelberg (2014)
7. Domingos, P., Richardson, M.: Mining the network value of customers. In: KDD, pp. 57–66 (2001)
8. Kempe, D., Kleinberg, J., Tardos, É.: Maximizing the spread of influence through a social network. In: KDD, pp. 137–146 (2003)
9. Johnson, A.: Nike-tops-list-of-most-viral-brands-on-facebook-twitter (2010). <http://www.kikabink.com/news/>
10. Hugo, O., Garnsey, E.: The emergence of electronic messaging and the growth of four entrepreneurial entrants. *New Technol. Based Firms New Millenium* **2**, 97–123 (2002)
11. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of approximations for maximizing submodular set functions—I. *Math. Program.* **14**(1), 265–294 (1978)
12. Saito, K., Nakano, R., Kimura, M.: Prediction of information diffusion probabilities for independent cascade model. In: Lovrek, I., Howlett, R.J., Jain, L.C. (eds.) KES 2008, Part III. LNCS (LNAI), vol. 5179, pp. 67–75. Springer, Heidelberg (2008)
13. Goyal, A., Bonchi, F., Lakshmanan, L.V.S.: Learning influence probabilities in social networks. In: WSDM, pp. 241–250 (2010)

14. Yang, J., Leskovec, J.: Modeling information diffusion in implicit networks. In: ICDM, pp. 599–608 (2010)
15. Gomez, R.M., Leskovec, J., Krause, A.: Inferring networks of diffusion and influence. In: KDD, pp. 1019–1028 (2010)
16. Durrett, R.: *Lecture Notes on Particle Systems and Percolation*. Wadsworth Publishing, Boston (1988)
17. Liggett, T.M.: *Interacting Particle Systems*. Springer, Heidelberg (1985)
18. Granovetter, M.: Threshold models of collective behavior. *Am. J. Sociol.* **83**(6), 1420–1443 (1978)
19. Schelling, T.: *Micromotives and Macrobehavior*. Norton, New York (1978)
20. Chen, W., Lakshmanan, L., Castillo, C.: *Information and Influence Propagation in Social Networks*. Morgan & Claypool, California (2013)
21. Horel, T., Singer, Y.: Scalable methods for adaptively seeding a social network. In: WWW, pp. 441–451 (2015)
22. Wang, C., Chen, W., Wang, Y.: Scalable influence maximization for independent cascade model in large-scale social networks. *Data Min. Knowl. Disc.* **25**(3), 545–576 (2012)
23. Chen, W., Lu, W., Zhang, N.: Time-critical influence maximization in social networks with time-delayed diffusion process. In: AAAI, pp. 1–5 (2012)
24. Liu, B., Cong, G., Zeng, Y., et al.: Influence spreading path and its application to the time constrained social influence maximization problem and beyond. *IEEE Trans. Knowl. Data Eng.* **26**(8), 1904–1917 (2014)
25. Wasserman, S., Faust, K.: *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge (1994)