Influence in Social Networks: A Unified Model?

Ajitesh Srivastava
Department of Computer Science
University of Southern California
Los Angeles, USA
ajiteshs@usc.edu

Charalampos Chelmis, Viktor K. Prasanna
Ming Hsieh Department of Electrical Engineering
University of Southern California
Los Angeles, USA
{chelmis, prasanna}@usc.edu

***Abstract*—Understanding how information flows in online social networks is of great importance. It is generally difficult to obtain accurate prediction results of cascades over such networks, therefore a variety of diffusion models have been proposed in the literature to simulate diffusion processes instead. We argue that such models require extensive simulation results to produce good estimates of future spreads. In this work, we take a complimentary approach. We present a generalized, analytical model of influence in social networks that captures social influence at various levels of granularity, ranging from pairwise influence, to local neighborhood, to the general population, and external events, therefore capturing the complex dynamics of human behavior. We demonstrate that our model can integrate a variety of diffusion models. Particularly, we show that commonly used diffusion models in social networks can be reduced to special cases of our model, by carefully defining their parameters. Our goal is to provide a closed-form expression to approximate the probability of infection for every node in an arbitrary, directed network at any time t. We quantitatively evaluate the approximation quality of our analytical solution as compared to numerous popular diffusion models on a real-world dataset and a series of synthetic graphs.**

***Keywords*—analytical framework; computational models; diffusion models; dynamics; evolutionary models; influence; statistical modeling; social networks; social simulation**

I. Introduction

With the proliferation of online social networks, researchers have tried to model, understand and make predictions of diffusion processes [1], [2]. Two of the most widely used diffusion models on social networks are the Linear Threshold Model (LTM) [3], and the Independent Cascade Model (ICM) [4]. It was recently shown that ICM and LTM are special cases of the Genetic Algorithm Diffusion Model (GADM), which emulates social interactions through a tail-swap cross-over interaction [2], assuming that social interactions are always pairwise. In our work, we propose a model to capture not only pairwise influence, but also local neighborhood effects, aggregate social behavior, and external factors, or a combination of them.

Typically proposed methodologies for influence calculation and models of diffusion need extensive simulation results to be evaluated, usually by means of statistical analysis. Instead, we devise a novel formulation of progressive diffusion with minimum computational complexity. It has been shown that exact computation of infection probabilities is #P-hard [5]. We provide an approximate generalized, analytical solution to the diffusion mechanism that comprises of two processes unfolding over the network simultaneously: (a) pairwise influence, and (b) pressure from collective dynamics, which can be a result of local social pressure, global influence, or external forces, or a combination of the above. Our methodology is vertex-centric, i.e., models each user separately, offering great flexibility in terms of modeling personalized influence functions, and allows for the use of time-dependent influence functions. Note, that in this work, we are not concerned with learning the parameters that drive the spread of infection from observational data. While this aspect is important, it is outside of the scope of this paper. To the best of our knowledge, our work is the first to (a) enable analytical computation of complex, non-linear phenomena like influence, while considering multiple factors that can change over time (Section II), (b) without requiring extensive simulation runs to estimate the propagation probabilities at the steady state. Our formula explicitly and formally unites a rich class of popular diffusion processes in social networks [6], [4], [1], [7] as special cases (Section III).

II. Analytical Model of Influence

We model the social network as a directed graph $G = (V, E)$, where a node $v \in V$ represents an individual, and edge $(v, u) \in E$ exists if v interacts with u (in our context v influences u). For every node v, we define the set of incoming neighbors $N_i(v) = \{u|(u, v) \in E\}$, and the set of outgoing neighbors $N_o(v) = \{u|(v, u) \in E\}$. Our goal is to model the probability of infection for every node in the network at any time t. Typically, in a diffusion process, a node can exist in one of two states at a given time - infected or susceptible. Here, we study the problem of progressive diffusion, where nodes that are infected do not become healthy again, i.e., they do not return to the susceptible state. Hence, every node can be infected once, and once infected stays infected.

A. Unified Model of Influence

We start with a seed set $S \subset V$ of infected nodes at time $t = 0$. The infection process proceeds in discrete time steps, in which two types of influence unfold over the network [7]. According to the first process, each infected node v attempts to infect its neighbors (*individual* influence). The probability of infection $p_{(v,u)}(t)$ is pairwise and may change over time. The second source of influence we consider is *collective* influence. According to this process, each susceptible node u can be infected with probability $r_u(t)$, independent of individual influence. This may include external factors [8], [9], [10], or external sources of exposure [11], or the status of the incoming neighborhood of u [7]. The function $r_u(t)$ is node specific, and may be time dependent.

B. Infection Probability Formula Under the Unified Model

Let $B_{u,t}$ represent the probability of infection of node u by the time t. Initial values $\{B_{u,0}\}$ are either 0 or 1 depending on the membership of u in the seed set. Let $E_{v,t}$ denote the indicator variable, which is 1 if node v is infected by the time t, 0 otherwise. To find the probability of a node u being infected at time t, we consider an arbitrary ordering of its incoming neighbor set $N_i(u)$: $< v_1, v_2, \ldots, v_n >$. Based on this, we define *zero state probability* at time $t-1$: $P^0_{s_n,s_{n-1},\ldots,s_1}$, where superscript 0 denotes $E_{u,t-1} = 0$. The subscript is a vector, which elements s_i denote the value of $E_{v_i,t-1}$, and can take values in $\{0, 1, *\}$. $s_i = 0$ represents $E_{v_i,t-1} = 0$, $s_i = 1$ denotes $E_{v_i,t-1} = 1$, and $s_i = *$ indicates marginalization over the state of v_i, i.e., '$E_{v_i,t-1} = 0$ or 1'. For instance, for a node u with four neighbors, $P^0_{0,1,*,1}$ denotes the probability $P(E_{u,t-1} = 0, E_{v_4,t-1} = 0, E_{v_3,t-1} = 1, E_{v_1,t-1} = 1)$. We begin by calculating $B_{u,t}$ in the special case of G being a tree, i.e., each node has at most one incoming neighbor.

Theorem 1: The infection probability of node u with parent v in a tree is given by:

$$B_{u,t} = 1 - (1 - r_u(t))\big((1 - p_{v,u}(t))(1 - B_{u,t-1}) + p_{v,u}(t)(1 - B_{v,t-1}) \prod_{k=1}^{t-1}(1 - r_u(k))\big) . \quad (1)$$

Proof: The probability of node u not being infected by time t is $P(E_{u,t} = 0) = 1 - B_{u,t}$. Either one of two things must have happened for u not to be infected by time t. First, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 0, E_{u,t-1} = 0)$ if and only if collective influence $r_u(t)$ failed to infect u at time t. Intuitively, when the parent of u was not infected at time$t-1$, the only chance for u to be infected at time t is through collective influence $r_u(t)$, with probability $1 - r_u(t)$. Second, state $E_{u,t} = 0$ was reached from state $(E_{v,t-1} = 1, E_{u,t-1} = 0)$, i.e., when the parent of u was infected at time $t-1$, if and only if collective influence was unsuccessful, and furthermore v failed to infect u. As the two processes are independent, this can happen with probability $(1 - r_u(t))(1 - p_{v,u}(t))$. It follows that,

$$\begin{aligned} 1 - B_{u,t} &= P^0_1(1 - r_u(t))(1 - p_{v,u}(t)) + P^0_0(1 - r_u(t)) \\ &= (1 - r_u(t))(P^0_1(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))((P^0_* - P^0_0)(1 - p_{v,u}(t)) + P^0_0) \\ &= (1 - r_u(t))(P^0_*(1 - p_{v,u}(t)) + P^0_0 p_{v,u}(t))) , \quad (2) \end{aligned}$$

where $P^0_* = P(E_{u,t-1} = 0) = 1 - B_{u,t-1}$ and $P^0_0 = P(E_{u,t-1} = 0, E_{v,t-1} = 0)$. This means v and u were both susceptible at time $t-1$. If v was also not infected by the time $t-1$, u can only be susceptible because all collective influence till that time failed, i.e., $P^0_0 = (1 - B_{v,t-1}) \prod_{k=0}^{t-1}(1 - r_u(k))$. We set $r_u(0) = 1$ if $u \in S$, 0 otherwise. Substituting the values of P^0_* and P^0_0 in Equation 2, results in Equation 1. ∎

Next, we extend Equation 1 to a graph of any type. Without loss of generality, we focus on directed graphs, as undirected graphs can be converted into their directed equivalent.

Lemma 1: The probability of a node u not being infected by the time t is related to the zero state probabilities as follows

$$1 - B_{u,t} = (1 - r_u(t)) \sum_{s_i \in \{0,*\}} \left(P^0_{s_n,s_{n-1},\ldots,s_1} \prod_{i=1}^{n} (1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{n} p_{v_i,u}(t)^{\delta_{s_i,0}} \right) . \quad (3)$$

where $\delta_{a,b} = 1$ only if $a = b$, 0 otherwise.

Proof: When the number of incoming neighbors is one, Lemma 1 follows from Equation 2. Now, suppose the statement is true for $k \geq 1$ parents. Consider a sequence $\mathbf{x_k} =< s_k, s_{k-1}, \ldots, s_1 >$. We look at the new terms that are added due to the inclusion of v_{k+1}. For ease of notation, let $D(\mathbf{x_k}) = (1 - r_u(t)) \prod_{i=1}^{k} (1 - p_{v_i,u}(t))^{\delta_{s_i,*}} \prod_{i=1}^{k} p_{v_i,u}(t)^{\delta_{s_i,0}}$. Equation 3 can be rewritten as

$$1 - B_{u,t} = \sum_{\mathbf{x_n}} P^0_{\mathbf{x_n}} D(\mathbf{x_n}) . \quad (4)$$

We have assumed that this is true for $n = k$. The addition of v_{k+1} affects $P(E_{u,t})$ in ways similar to those discussed in Theorem 1, i.e., if $E_{v_{k+1},t-1} = 1$, then this new node fails to infect u with probability $(1 - p_{v_{k+1},u}(t))$. On the other hand, if $E_{v_{k+1},t-1} = 0$, node v_{k+1} does not have the ability to infect. Formally, the new terms added are:

$$\begin{aligned} & P^0_{1,\mathbf{x_k}}(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}} D(\mathbf{x_k}) \\ =& (P^0_{*,\mathbf{x_k}} - P^0_{0,\mathbf{x_k}})(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}} D(\mathbf{x_k}) \\ =& P^0_{*,\mathbf{x_k}}(1 - p_{v_{k+1},u}(t))D(\mathbf{x_k}) + P^0_{0,\mathbf{x_k}} p_{v_{k+1},u}(t) D(\mathbf{x_k}) \\ =& P^0_{*,\mathbf{x_k}} D(*, \mathbf{x_k}) + P^0_{0,\mathbf{x_k}} D(0, \mathbf{x_k}) , \end{aligned}$$

which would generate the required terms in the right hand side of Equation 4, when $n = k + 1$. This indicates that the statement is true for $k + 1$ incoming neighbors. By induction, Lemma 1 is true $\forall n$. ∎

Theorem 2: An approximate probability of infection is given by the recurrence relation:

$$\begin{aligned} B_{u,t} = 1 - \Bigg[& (1 - B_{u,t-1}) \Big(\prod_{v \in N_i(u)} (1 - p_{v,u}(t) B_{v,t-1}) \Big) \\ & + \Big(\prod_{v \in N_i(u)} p_{v,u}(t)(1 - B_{v,t-1}) \Big) \\ & \Big(\prod_{k=1}^{t-1} (1 - r_u(k)) - 1 + B_{u,t-1} \Big) \Bigg] (1 - r_u(t)) . \quad (5) \end{aligned}$$

The approximation comes from assuming that the states of infection of incoming neighbors of a given node u are independent, i.e., for two incoming neighbors v_i and v_j, events $E_{v_i,t-1} = 0$ and $E_{v_j,t-1} = 0$ are independent. Next, we proceed with proving the Theorem.

Proof: We attempt to find the zero state probabilities for sequence $\mathbf{x_n}$. When $\mathbf{x_n} =< 0, 0, \ldots, 0 >$, u and all nodes in $N_i(u)$ are susceptible, which means that collective influence till $t-1$ was unsuccessful. Further, at $k = 0$, $r_u(t) = 1$ for $u \in S$. In this case,

$$P^0_{0,0,\ldots,0} = \prod_{k=0}^{t-1} (1 - r_u(k)) \prod_{v \in N_i(u)} (1 - B_{v,t-1}) . \quad (6)$$

Any other sequence $\mathbf{x_n}$, which consists of at least one $*$ in the i-th position, represents the state of u being not infected by the state of its i-th neighbor. Given the state of u's neighbors, the conditional probability of u not being infected is $1 - B_{u,t-1}$. The zero state probability is then computed as follows

$$P^0_{\mathbf{x_n}} = (1 - B_{u,t-1}) \prod_{s_i=0} (1 - B_{v_i,t-1}) \,. \tag{7}$$

Combining Equations 4, 6 and 7, and some algebraic manipulations lead to Equation 5, thus completing the proof. ■

C. Complexity Analysis

The recurrence relation in Equation 5 requires inspection of all incoming links to a node u, $|N_i(u)|$, at every time step. Therefore, in order to evaluate Equation 5 for all nodes for t time steps, the number of operations required is $t \sum_u O(|N_i(u)|) = O(|E|t)$.

III. Reduction to Other Models

A. Complex Contagion Model

According to the Complex Contagion Model [7], infection can be achieved at time t in two ways. First, each node that was infected at time $t-1$ attempts to infect each of its outgoing neighbors with probability p. Once all infected nodes are examined, healthy nodes have a chance of random infection. For n_i^{t-1} infected nodes by the time $t-1$, the probability of random infection at time t is given by an exponential growth law: $r(t) = \exp(\alpha n_i^{t-1} - \beta)$, where α and β are constants [7].

Reduction: We model individual influence as

$$p_{v,u}(t) = p, \forall v, u, t \,. \tag{8}$$

Substituting collective influence with the random infection factor results in

$$r_u(t) = r(t) = \exp(\alpha \sum_u B_{u,t-1} - \beta) \,, \tag{9}$$

since the number of infections by the time $t-1$, n_i^{t-1}, is computed as follows:

$$n_i^{t-1} = \mathbb{E}(\sum_u E_{u,t-1}) = \sum_u \mathbb{E}(E_{u,t-1}) = \sum_u B_{u,t-1} \,. \tag{10}$$

Equations 8 and 9 form the reduction of Unified Model to the Complex Contagion Model. ■

B. Independent Cascade Model

In the Independent Cascade Model [4], a seed set of infected nodes is provided. At each time step t, each node is either infected or susceptible. Every node v that was infected at time $t-1$ has a single chance to infect each of its neighbors u. The infection succeeds with probability $p_{v,u} = p$ [4].

Reduction: At any time t, a susceptible node u has a single chance to be infected by its neighbors that were infected at $t-1$. If at least one of them succeeds, u gets infected. The probability of node u getting infected is then given by

$$\begin{aligned} r_u(t) &= P(\text{at least one infected neighbor succeeds}) \\ &= 1 - P(\text{no infected neighbor succeeds}) \\ &= 1 - \prod_{v \in N_i(u)} (1 - p(A_{v,t-1})) \,, \end{aligned} \tag{11}$$

where $A_{v,t-1}$ is the probability of v being infected at time $t-1$. Since $B_v^t = \sum_{\tau=0}^{t} A_{v,\tau}$, if follows that:

$$A_{v,t} = \begin{cases} B_{v,t-1} & \text{if } t = 1 \\ B_{v,t-1} - B_{v,t-2} & \text{if } t > 1 \end{cases}$$

The individual influences are set to zero. This concludes the reduction. ■

C. Threshold Models

In threshold models the probability of infection of a node depends on the popularity of the contagion in its incoming neighborhood. Several threshold models exist in the literature, including the Linear Threshold Model [3], and the Linear Friendship Model [1].The Generalized Threshold Model [4] dictates that a node u is infected based on a monotone function of the set of its infected neighbors $f(In(u,t)) \in [0,1]$ and a threshold $\theta_u \in [0,1]$. Particularly, u is infected at time t if $f(In(u,t)) \geq \theta_u$. Note that the the threshold θ_u can be randomly selected at each time t [12] leading to non-determinism of the infection process. Since, these thresholds are selected uniformly at random, this is equivalent to saying that the probability of infection of a healthy node u at time t is $f(In(u,t))$.

Reduction: At any time t, the probability of node u getting infected is given by a function of u's status as follows:

$$P(u \text{ infected at time } t) = f(In(u,t), \vec{b_u}) \,, \tag{12}$$

where $In(u,t) = \{v | v \in N_i(u), E_{v,t-1} = 1\}$ and $\vec{b_u} = \{b_{v,u} | v \in N_i(u)\}$ is a vector of pairwise weights $b_{v,u}$ associated to v's incoming neighbors. Substituting $r_u(t)$ in Equation 5 with Equation 12, and setting all individual influence probabilities to zero, concludes the reduction. ■

For instance, the Linear Friendship Model (LFM) [1] the probability of node u getting infected at time t is given by

$$P(u \text{ infected at time } t) = \frac{\exp(\alpha |In(u,t)| + \beta)}{1 + \exp(\alpha |In(u,t)| + \beta)}, \tag{13}$$

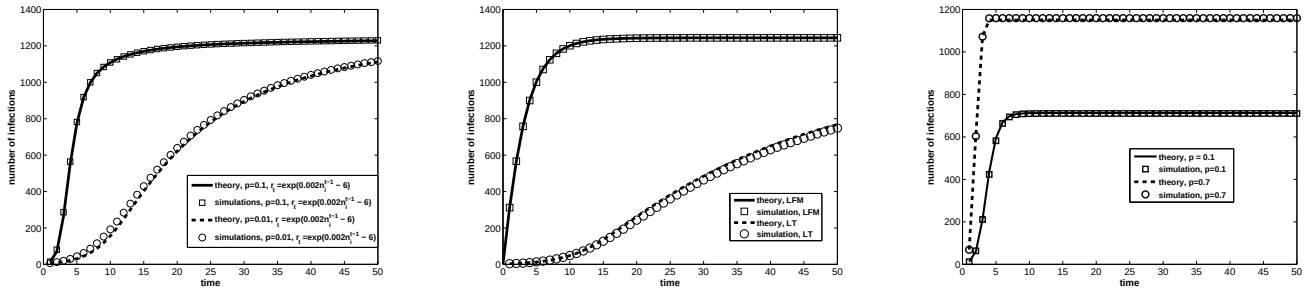
where $|In(u,t)| = \sum_{v \in N_i(u)} B_{v,t-1}$, i.e., the number of infections by the time $t-1$ is calculated similarly to the Complex Contagion Model [7] using Equation 10, with the difference that the population is restricted to the local neighborhood of node u. Including pairwise weights $b_{v,u}$ into the formulation results in $|In(u,t,\vec{b_u})| = \sum_{v \in N_i(u)} b_{v,u} B_{v,t-1}$, which concludes the reduction.

IV. Experiments

With our model well-defined, we now apply it to a real life dataset from popular social news aggregator Digg[1], and a series of synthetic data. We used a subset of Digg's[2] follower graph. Our dataset consists of $1,244$ nodes and $28,343$ directed links. A link from v to u exists if v influences u, i.e., when u follows v. Table I summarizes the set of parameters used in our experiments. For each model, we start with a seed set of two infected nodes. Figure 1 shows the results of infection spreads over time using average of 1000 simulations for each model,

[1] http://digg.com/

[2] Dataset url: http://www.public.asu.edu/~ylin56/kdd09sup.html



(a) Complex Contagion Model (b) Threshold Models (c) Independent Cascade Model

Fig. 1. Agreement of simulation and theory for the three models for Digg1k dataset.

TABLE I. PARAMETERS USED IN THE EXPERIMENTAL VALIDATION ON DIGG FOLLOWER GRAPH

parameter set 1	CCM	$p = 0.1, r(t) = exp(0.002n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \vec{b_u}) = \sum_{v \in In(u,t)} b_{v,u}$
	ICM	$p = 0.1$
parameter set 2	CCM	$p = 0.01, r(t) = exp(0.002n_i^{t-1} - 6)$
	GLT	$f(In(u,t), \vec{b_u}) = \frac{\exp(\lvert In(u,t,\vec{b_u})\rvert)}{1+\exp(\lvert In(u,t,\vec{b_u})\rvert)}$
	ICM	$p = 0.7$

and the corresponding predicted values obtained by using the analytical solution from Theorem 2. The prediction matches very well with the average simulations, providing an empirical, quantitative confirmation that Equation 5 produces a good fit to the expected outcome which is obtained by computationally expensive simulation runs. For an extensive analysis of the approximation error, we ran a series of experiments on synthetic graphs with varying sizes and densities. The figures have been omitted for brevity. To summarize, we found that the error remains small for different graph sizes and densities. The error is also unaffected by the various models' parameter values. This fact empirically verifies our claim that under the Unified Model of Influence, Equation 5 is a good approximation to various models.

V. CONCLUSION

Influence analytics and diffusion prediction in online social networks have been important for many domains from marketing to public health. With the tremendous increase in the volume of data, network sizes reach millions of nodes, restricting the applicability of computational models for diffusion prediction. In this work, we have proposed a novel, general analytical framework for influence calculation in social networks, which does not require extensive simulations. In this framework, each node has its own individual function of collective influence and pairwise influence functions for each neighboring node. Both functions vary with time, thus making our framework directly applicable to a plethora of situations. We have shown how various popular models of diffusion constitute special cases of our model. Hence, our formula is applicable for approximating the expected outcome of these models. We have validated our results using a real world social network and a number of synthetic graphs.

ACKNOWLEDGMENT

This work is supported by Chevron Corp. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

REFERENCES

[1] C. Budak, D. Agrawal, and A. El Abbadi, "Diffusion of information in social networks: Is it all local?" in *2012 IEEE 12th International Conference on Data Mining (ICDM)*, 2012, pp. 121–130.

[2] A. Hajibagheri, A. Hamzeh, and G. Sukthankar, "Modeling information diffusion and community membership using stochastic optimization," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 175–182.

[3] M. Granovetter, "Threshold models of collective behavior," *American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[4] D. Kempe, J. Kleinberg, and E. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2003, pp. 137–146.

[5] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 1029–1038.

[6] T. W. Valente, "Social network thresholds in the diffusion of innovations," *Social Networks*, vol. 18, no. 1, pp. 69–89, 1996.

[7] C. Chelmis and V. K. Prasanna, "The role of organization hierarchy in technology adoption at the workplace," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13. New York, NY, USA: ACM, 2013, pp. 8–15.

[8] E. Abrahamson and L. Rosenkopf, "Social network effects on the extent of innovation diffusion: A computer simulation," *Organization Science*, vol. 8, no. 3, pp. 289–309, 1997.

[9] F. M. Bass, "A new product growth for model consumer durables," *Manage. Sci.*, vol. 50, no. 12 Supplement, pp. 1825–1832, December 2004.

[10] H. Choi, S.-H. Kim, and J. Lee, "Role of network structure and network effects in diffusion of innovations," *Industrial Marketing Management*, vol. 39, no. 1, pp. 170 – 177, 2010.

[11] S. A. Myers, C. Zhu, and J. Leskovec, "Information diffusion and external influence in networks," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '12. New York, NY, USA: ACM, 2012, pp. 33–41.

[12] Z. Lu, W. Zhang, W. Wu, J. Kim, and B. Fu, "The complexity of influence maximization problem in the deterministic linear threshold model," *Journal of combinatorial optimization*, vol. 24, no. 3, pp. 374–378, 2012.