Adaptive Influence Maximization in Dynamic Social Networks

Guangmo Tong, *Student Member, IEEE*, Weili Wu, *Member, IEEE*,
Shaojie Tang, and Ding-Zhu Du, *Member, IEEE*

***Abstract*— For the purpose of propagating information and ideas through a social network, a seeding strategy aims to find a small set of seed users that are able to maximize the spread of the influence, which is termed influence maximization problem. Despite a large number of works have studied this problem, the existing seeding strategies are limited to the models that cannot fully capture the characteristics of real-world social networks. In fact, due to high-speed data transmission and large population of participants, the diffusion processes in real-world social networks have many aspects of uncertainness. As shown in the experiments, when taking such uncertainness into account, the state-of-the-art seeding strategies are pessimistic as they fail to trace the influence diffusion. In this paper, we study the strategies that select seed users in an adaptive manner. We first formally model the dynamic independent Cascade model and introduce the concept of adaptive seeding strategy. Then, based on the proposed model, we show that a simple greedy adaptive seeding strategy finds an effective solution with a provable performance guarantee. Besides the greedy algorithm, an efficient heuristic algorithm is provided for better scalability. Extensive experiments have been performed on both the real-world networks and synthetic power-law networks. The results herein demonstrate the superiority of the adaptive seeding strategies to other baseline methods.**

***Index Terms*— Social network influence, adaptive seeding strategy, stochastic submodular maximization.**

I. Introduction

WITH the advancements in information science in the last two decades, social networks are becoming important dissemination platforms as they allow efficient interchange of ideas and information. The process of influence diffusion in social networks has been studied in many domains e.g. epidemiology, social median and economics. It has been shown that the investigation into the influence diffusion are of great use in many aspects such as designing marketing strategy [1], [2], analyzing human behavior [3] and rumor blocking [4]. In order to formulate the diffusion process, a number of models have been studied during the last decade. Two basic operational models, linear threshold (LT) model and independent cascade (IC) model, are proposed by Kempe *et al.* [5]. In the Linear Threshold Model, a user will adopt a new idea if the influence from its neighbors has reached a certain threshold, while in the Independent Cascade Model an adopter has a certain probability to convince each of its neighbors. Based on those basic models various advanced models have been developed and studied.

Among the topics regarding influence diffusion, an important one is that how to propagate information through a social network effectively and efficiently. As an example, in order to advertise new products, a company would like to offer free samples to a set of initial users who will potentially introduce the new product to their friends. Due to expense issue, only a limited number of samples are available and thus we have a budget of the seed users. A natural problem is that how to select a good set of seed users that is able to maximize the number of customers who finally adopt the target product. This problem is named as influence maximization problem first proposed in [6] in literature.

Although a large body of related works have been performed concerning the influence maximization problem, the state-of-art technique may not be able to deal with many real cases in effect. A drawback of the existing diffusion models is that they fail to take account of some uncertain natures of the diffusion process in a real-world social network. Such uncertainness can be viewed from the following aspects. In a real-world social network, the seed users are not assured to be successfully activated. In the example of selling a new product, the advertising would be stuck if the free samples do not satisfy the initial users. Furthermore, due to frequent variation of the degree of the relationship between users, the topology of a social network is not always static in real cases. In the sense of an online social network, such as Facebook, Twitter or Flicker, topology changes are incurred by the increasing number of the common friends between a pair of users. In this paper, we study the influence maximization problem in the social networks with the above characteristics. By extending the classic IC model, we herein develop the Dynamic Independent Cascade (DIC) model which is able to better capture the dynamic aspects of real social networks. In the classic IC model a seed node is guaranteed to be activated after selected and the relationship between two users is simply represented by a fixed probability, while the seed nodes in our DIC model could fail to be activated with a certain probability and the propagation probability between two users

Manuscript received July 1, 2015; revised January 12, 2016 and March 29, 2016; accepted April 24, 2016; approved by IEEE/ACM TRANSACTIONS ON NETWORKING Editor Y. Yi.

G. Tong, W. Wu, and D.-Z. Du are with the Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: gxt140030@utdallas.edu; weiliwu@utdallas.edu; dzdu@utdallas.edu).

S. Tang is with the Naveen Jindal School of Management, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: shaojie.tang@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNET.2016.2563397

follows a certain distribution[1] which reflects the change of topology of a social network.

Based on the DIC model, we further consider how to design a seeding strategy that can find effective seed nodes. For the classic IC model, Kempe *et al.* [5] propose a simple greedy algorithm with an approximation ratio of $(1 - 1/e)$ and Chen *et al.* [8] present an efficient heuristic seeding approach to handle large-scale social networks. The existing approaches always make seeding selection in a static manner (i.e., determining a seed set before the start of spread), which renders them inapplicable to the DIC model. As mentioned earlier the seed users in the DIC model are not guaranteed to be activated. In this setting, an arising issue is that we can seed a user for more than one time if it is not successfully activated in the past rounds. One can see that it is worthy to take more effort to activate a powerful user as he or she may generate considerable influence to a social network. However, a static seeding algorithm cannot take such a case into account. Besides, to determine a seed set, the prior algorithms require the propagation probability between users, but in the DIC model such a probability is a random variable and we can only expect a distribution over it. Admittedly we could take advantage of its expected value and then apply the existing approaches, but such a method would be pessimistic as it fails to trace the dynamic topology of a real-world social network. In order to better support the dynamic social networks, we herein study the adaptive seeding strategies which make seeding decisions step by step according to the observed influence diffusion.

A. Related Work and Technique

Domingos and Richardson [6] are among the first who study the influential nodes in viral marketing. In the seminal work [5], Kempe *et al.* formulate the influence maximization problem from the view of combinatorial optimization, and provide a greedy algorithm with an approximation ratio of $(1 - 1/e)$ in LT and IC models. Efficient heuristic influence maximization algorithms have been studied in many works e.g. [8]–[10]. Long and Wong [11] further study this problem from the perspective of minimization. Du *et al.* [12] and Gomez-Rodriguez and Schölkopf [13] propose the continuous diffusion model and study the influence maximization problem in this setting. All the above works aim to determine an effective seed set before the diffusion process and focus on the network with a static topology.

In order to learn a provable performance guarantee, submodular function plays an important role in the prior works. Kempe *et al.* [5] show that the expected number of active nodes is a monotonically increasing submodular function over the seed set, and therefore, by the celebrated result in [14], a simple greedy algorithm yields an $(1-1/e)$-approximation. However, as shown later in Sec. III, such a technique cannot be directly applied to the adaptive seeding problem. On the one hand the seed nodes are unknown before the diffusion process as they are adaptively selected; on the other hand the value of the objective function over a certain seed set cannot be explicitly observed.

Adaptive seeding strategy is a stochastic optimization framework and a natural extension to original seeding approach in [5]. Part of the analysis in this paper is based on the stochastic submodular maximization. Asadpour *et al.* [15] present the analysis of the original stochastic submodular maximization problem where the objective function is defined on the power set of a set of independent random variables. Golovin and Krause [16] further study this problem with the concept of adaptive submodularity. Although the above works are only applicable to some special cases of the adaptive influence maximization problem considered in this paper, they provide a clue that the greedy algorithm in its adaptive version is still able to achieve a provable performance guarantee. In recent works, Seeman and Singer [17] consider the adaptive approach to a variant influence maximization problem, where the seed nodes are constrained in a certain set and the influence can spread for only one round, and Horel and Singer in [18] consider the strategies with two seeding steps. In this paper, we will focus on general adaptive seeding strategies.

B. Contribution

The contributions of this paper are summarized as follows.

1) We propose the DIC model that is able to capture the dynamic aspects of real-world social networks, and formally define the adaptive seeding strategy with the concept of seeding pattern.
2) We construct an optimal seeding pattern and propose an adaptive hill-climbing strategy with a provable performance guarantee in the DIC model.
3) Inspired by a crucial observation in simulations, we further design an efficient heuristic adaptive seeding strategy which narrows the candidate seed set before the seeding process.
4) The conducted experiments demonstrate the superiority of the proposed adaptive seeding strategies to the existing seeding approaches in dynamic social networks.

The rest of the paper is organized as follows. The proposed DIC model and the adaptive seeding strategy are formulated in Sec. II. The analysis of the greedy adaptive strategy is shown in Sec. III and the heuristic strategy is proposed in Sec. IV. In Sec. V, we show the experimental results. Sec. VI concludes.

II. Problem Setting

A. DIC Model

A social network is modeled as a directed graph where nodes and edges denote the individuals and social ties, respectively. In order to spread an idea or advertising a new product in a social network, some seed nodes are chosen to be activated (e.g., by giving payments or offering free samples) to trigger the spread of influence. Inherit from [5] we speak of each node as being either *active* or *inactive*. Initially, all the nodes are inactive. A node can be activated either by its neighbor or as

[1] It can be inferred from the strong ties [2], [7] or other elements that affect the affinity of users in a social network.

TABLE I
NOTATIONS

Symbol	Definition
G	Instance of DIC network.
G_1	Example DIC network in Example 1.
B	Budget of seed set.
D_e	Domain of the propagation probability of edge e.
d_e^i	The i^{th} value in D_e.
Prob$[X_u = 1]$	The probability that u can be activated as a seed node when selected.
A	Seeding pattern.
A_0	Special seeding pattern define in Def. 3.
A^*	Special seeding pattern define in Def. 4.
S_A^G	Seeding strategy of pattern A on G
OPT_A^G	Optimal seeding strategy of pattern A on G
c-G	Auxiliary graph of network G
x	Full realization
$S_A^{G_x}$	Seeding strategy of pattern A on G under f-realization x
y	Partial realization
ϵ	Empty realization

a seed node. The diffusion process goes round by round: in round i, every active node remains active and it activates its neighbors with a certain probability.

In the DIC model, associated with each node u there is a random variable X_u following a Bernoulli distribution f_u, where $X_u = 1$ indicates node u is successfully activated as a seed node. For the relationship between nodes, an active node u has one chance to activate its inactive neighbor v via edge (u, v) with a propagation probability of $X_{(u,v)}$ which is a random variable. As discussed in [19] and [20], the propagation probability varies from time to time and it is hard to be estimated without prior knowledge. Therefore, assuming a distribution, instead of a fixed value, for propagation probability is more reasonable for real social networks. Without loss of generality, for each edge e, we assume that X_e follows a certain discrete distribution f_e with a domain D_e, and let $d_e^i \in [0, 1]$ be the i^{th} value in D_e. In this paper, we do not enforce any specific distribution of X_e.[2] In the DIC model, for an edge $e = (u, v)$, the value of X_e remains unknown until one of the neighbors of u is active. This is because in practice an industry institute may only trace the interested influence and the real-time state of the rest of the network is unavailable. We denote an instance of DIC network by $G = (V, E, F_V, F_E)$, where $F_V = \{f_u | u \in V\}$ and $F_E = \{f_e | e \in E\}$ are the sets of the distributions of X_u and X_e, respectively. Let N be the number of the nodes in V. Due to the expense of activating seed nodes, there is a budget $B (B \leq N)$ for the seed set. The notations that are frequently used later in this paper are listed in Table I and the rest of the notations in Table I will be introduced later.

B. Adaptive Seeding Strategy

Assuming that the seed nodes are only selected between two spread rounds, we denote the seeding step between round $i-1$ and round i as the i^{th} seeding step, and the first seeding step is executed before the process of spread. We assume that we need one round to activate the seed nodes selected in each seeding step. In this paper, we preserve *step* for seeding process and *round* for diffusion process.

Basically, to design an adaptive seeding strategy we consider two problems: (1) how many budgets should we use in each seeding step and (2) which nodes to select. We employ the following concepts to formulate those problems.

Definition 1: A *seeding pattern* $A = (a_1, \ldots, a_N)$ is a sequence of non-negative integers, implying that we seed a_i nodes in the i^{th} seeding step. We will later show that it suffices to consider a pattern that has at most N seeding steps. Due to budget constraint, $\sum a_i \leq B$. Note that it reduces to a non-adaptive seeding strategy if $A = (B)$. Corresponding to a seeding pattern $A = (a_1, \ldots, a_N)$, a *seeding strategy* $S_A = (s_1, \ldots, s_N)$ of A is a sequence of node-sets where $s_i \in V$ is the node-set seeded in the i^{th} seeding step and $|s_i| = a_i$. That $a_i = 0$ implies that we do not seed any node in the i^{th} seeding step and consequently $s_i = \emptyset$.

In the above setting, both the seeding pattern and seeding strategy can be adaptively constructed. That is, a_i and s_i may depend on the outcomes of the past rounds. For a specific DIC network G, we use S_A^G to denote a seeding strategy of pattern A on G. Since DIC model is a probabilistic model, the objective function herein is the expected number of the final active nodes when there is no node can be further activated and no budget left. We denote the expected number of active nodes in G under a seeding strategy S_A^G by $E[S_A^G]$.

Definition 2: Given a strategy S_A^G on a DIC network G, if, for some i, $s_i = \emptyset$ but there does not exist any edge (u, v) such that u is activated, either by its neighbors or as a seed node, in the $(i-1)^{th}$ round, we say that S_A^G waits for a *null round*. It can be easily seen that waiting for a null round has no impact on the process of spread or the effect of the strategy. Unless otherwise stated, we assume that any strategy will not wait for one or more null rounds. Therefore, there are at most N seeding steps and $s_1 \neq \emptyset$ for any strategy $S_A^G = (s_1, \ldots, s_N)$. For the convenience of analysis, we require that any strategy S_A^G will not select an active node as a seed node.

Two natural patterns A_0 and A^* are defined as follows.

Definition 3: Let $A_0 = (a_1, \ldots, a_N)$ where $a_i = 1$ for $1 \leq i \ \leq B$ and $a_i = 0$ for $i > B$. Informally, under pattern A_0 we successively seed one node in each step until the budget is used up.

Definition 4: Another pattern A^* is adaptively constructed as follows. In pattern A^*, we seed one node at a time and *wait until no node can be further activated* before seeding the next node. Thus, we seed one node in the first step and the rest of seeding pattern will be constructed adaptively. Note that under pattern A_0 we must have one node seeded in each of the first B seeding steps while under A^* we may not seed any node in a certain seeding step even there is still some budget available.

[2] We may assume an exponential distribution as a social network always exhibits a power-law pattern where the influential users are rare [21].

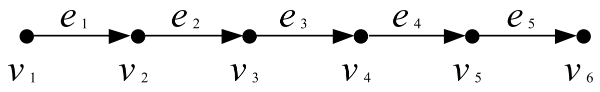


Fig. 1. Example network G_1.

Note that given a pattern A there exists many strategies of A. We use OPT_A^G to denote the optimal adaptive strategy of pattern A on a given DIC network G, with respect to the expected number of active nodes.

The core problem considered in this paper is defined as follows.

Problem 1 (Adaptive Influence Maximization (AIM)): Under the budget constraint, for any DIC network G, find a pattern A and a strategy S_A^G of A on G such that $E[S_A^G]$ is maximized.

C. An Example

We employ the following example to illustrate the DIC model and the concept of seeding pattern.

Example 1: Consider an example DIC network $G_1 = (V, E, F_V, F_E)$ with six nodes and five edges, as shown in Fig. 1, where $f_v(1) = 0.5$ for each $v \in V$, and $D_e = \{0.4, 0.8\}$ with $f_e(0.4) = 0.8$ for each $e \in E$. In this example, each node can be activated with a probability of 0.5 when selected as a seed node, and the propagation probability between two connected nodes could be 0.4 or 0.8 with probabilities 0.8 and 0.2, respectively. We set the budget B to be three. Suppose a certain seeding strategy $S_{A_1}^{G_1}$ produces a sequence of seed sets as $(\{v_3\}, \{v_3\}, \emptyset, \{v_1\})$ of pattern $A_1 = (1, 1, 0, 1)$. In this concrete seeding process, $S_{A_1}^{G_1}$ seeds v_3 twice respectively in step 1 and 2, which implies it fails to activate v_3 in the first time.

III. Greedy Algorithm

In this section, we show the main result of this paper. The seed selection rule of the greedy algorithm is shown as follows.

Rule 1: In each seeding step, we select the node that is able to maximize the marginal profit[3] conditioned on the observed events.

We will formally describe this algorithm later in Sec. III-C. Note that in each step we can observe the followings: (1) the influence diffusion of the past rounds; (2) the propagation probabilities between the active nodes and their neighbors. For a pattern A and a DIC network G, we use $\overline{S}_A^G$ to denote the seeding strategy following Rule 1. Our analysis consists of three steps. First, we propose a transformation approach which finds an explicit expression of the expected number of the active nodes. Then, we prove that A^* is the optimal pattern for any DIC network G, i.e., for any pattern A', $E[OPT_{A^*}^G] \geq E[OPT_{A'}^G]$. Finally, we show that $\overline{S}_{A^*}^G$ is an $(1 - 1/e)$-approximation under pattern A^*, i.e.,

$$E[\overline{S}_{A^*}^G] \geq (1 - 1/e) \cdot E[OPT_{A^*}^G].$$

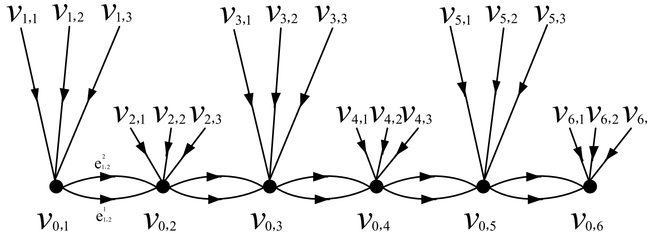


Fig. 2. Auxiliary graph c-G_1. In Example 1, we have a budget of three and the propagation probability of each edge in G_1 follows a distribution on a domain of two values. Therefore, there are three nodes $v_{1,1}$, $v_{1,2}$ and $v_{1,3}$ connected to $v_{0,1}$, and two edges $e_{1,2}^1$ and $e_{1,2}^2$ connecting $v_{0,1}$ and $v_{0,2}$.

A. Transformation

For a classic IC network, a concrete network is a graph where each edge (u, v) is specified to be either *live* or *not live*. If edge (u, v) is live then it means u could successfully activate v. Informally speaking, all the uncertainties are determinate in a concrete network. In a concrete network, the active nodes are those which are connected to a seed node via a path of live edges, and the number of the active nodes in a concrete network is a submodular function over the seed set [5]. Unfortunately, this approach cannot be directly applied to the analysis of our DIC model because several cases in the DIC model cannot be represented by a graph with a structure identical to that of the original DIC network. For example, how to represent the case that we seed a node more than once, and how to depict the feature that each propagation probability follows a distribution instead of being a single value? To address such scenarios, we transfer the original network to an auxiliary graph[4] where the active nodes can be explicitly observed given a seed set.

Given a DIC network $G = (V, E, F_V, F_E)$ where $V = \{v_1, \ldots, v_N\}$, we construct an auxiliary graph c-$G = (V_c, E_c)$, as follows. V_c consists of $N \cdot B + N$ nodes and is partitioned into $N + 1$ subsets denoted by V_c^i $(0 \leq i \ \leq N)$, where $|V_c^0| = N$ and $|V_c^i| = B$ $(i > 0)$. Let $V_c^0 = \{v_{0,1}, \ldots, v_{0,N}\}$ and $V_c^i = \{v_{i,1}, \ldots, v_{i,B}\}$ $(i > 0)$. Nodes in V_c^0 are corresponding to the nodes in G and nodes in V_c^i $(i > 0)$ are used to represent the multiple seedings on v_i in G. E_c consists of two parts E_c^1 and E_c^2, defined as follows. For $i > 0$ and $1 \leq j \ \leq B$, there is an edge $(v_{i,j}, v_{0,i})$ for each pair of $v_{i,j}$ and $v_{0,i}$, and for each pair of nodes $v_{0,i}$ and $v_{0,j}$ in V_0 $(1 \leq i \ \neq j \ \leq N)$, there are $|D_{(v_i,v_j)}|$ edges denoted by $e_{i,j}^k$ $(1 \leq k \ \leq |D_{(v_i,v_j)}|)$ connecting $v_{0,i}$ to $v_{0,j}$. Let E_c^1 be the set of edges between V_c^0 and V_c^i $(i > 0)$ and E_c^2 be the set of edges within V_0^i. Recall that $D_{(v_i,v_j)}$ is the domain of $f_{(v_i,v_j)}$ which is the distribution of the propagation probability of edge (v_i, v_j) in G.

The auxiliary graph c-G_1 of G_1 in Example 1 is illustrated in Fig. 2. Further explanations are presented in the caption.

Now we show that given a seeding strategy how to observe the active nodes via c-G. Following the notations in [15], we introduce the states of edges and the concept of realization.

[3] The mathematical definition will be given later in Eq. (5).

[4] The auxiliary graph of a DIC network is analogous to the concrete network of a classic IC network.

TABLE II
STATES OF EDGE $(v_{i,j}, v_{0,i})$ IN E_c^1, FOR $i > 0$ AND $1 \leq j \ \leq B$

State	**Explanation**
live	v_i in G is successfully activated when selected as a seed node in the j^{th} time.
not live	v_i in G fails to be activated when selected as a seed node in the j^{th} time.
undetermined	The result of the j^{th} seeding on v_i is unknown.

TABLE III
STATES OF $e_{i,j}^k$ IN E_c^2, FOR $1 \leq i \ \leq N$, $1 \leq j \ \leq N$ AND $1 \leq k \ \leq |D_{(v_i,v_j)}|$

State	**Explanation**
selected-live	The propagation probability between v_i and v_j is $d_{(v_i,v_j)}^k$ and v_i activates v_j.
selected-not live	The propagation probability between v_i and v_j is $d_{(v_i,v_j)}^k$ and v_i fails to activate v_j.
selected-undetermined	The propagation probability between v_i and v_j is $d_{(v_i,v_j)}^k$ and the result of the activation from v_i to v_j is unknown.
not selected	The propagation probability between v_i and v_j is not $d_{(v_i,v_j)}^k$.
undetermined	The propagation probability between v_i and v_j is unknown

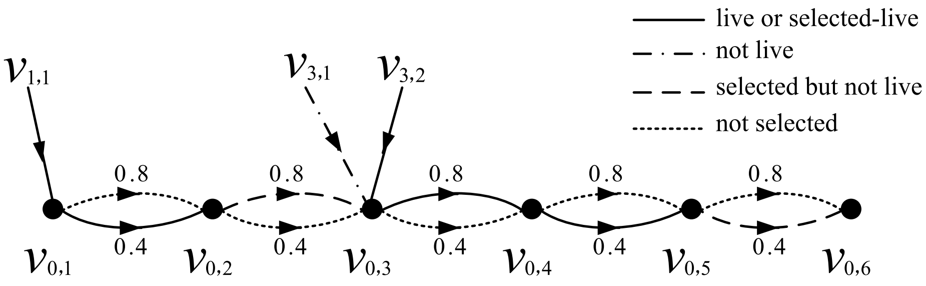


Fig. 3. An example f-realization x_1 of c-G_1. The number aligned with an edge is the propagation probability it stands for. In this concrete case, the seed nodes are v_1 and v_3, and the active nodes in G are v_1, v_3, v_4 and v_5.

Definition 5: A *full realization* (*f-realization*) x of c-G is a mapping from edges in c-G to some states, where each edge in E_c^1 is mapped to {*live*, *not live*} and each edge in E_c^2 is mapped to {*selected-live*, *selected-not live*, *not selected*}. In an f-realization, only one edge from $v_{0,i}$ to $v_{0,j}$ can be mapped to *selected-live* or *selected-not live*.

Definition 6: A *partial realization* (*p-realization*) y of c-G is a mapping from edges to states, where each edge in E_c^1 is mapped to {*live*, *not live*, *undetermined*}, and each edge in E_c^2 is mapped to {*selected live*, *selected-not live*, *not selected*, *selected-undetermined*, *undetermined*}. In a p-realization, if one edge from $v_{0,i}$ to $v_{0,j}$ is *undetermined* then all the edges from $v_{0,i}$ to $v_{0,j}$ must be *undetermined*; if one edge from $v_{0,i}$ to $v_{0,j}$ is either *selected-live*, *selected-live* or *selected-undetermined*, then others edges from $v_{0,i}$ to $v_{0,j}$ must be *not selected*.

The explanations of the states are listed in Tables II and III. Each edge together with its state in c-G corresponds to an event in the diffusion process of the original network G. We can see that an f-realization is a determinate case of the diffusion process and a p-realization is an intermediate state where the events are partially determined. With the concept of f-realization, for a seeding strategy S_A^G, the seed nodes selected by S_A^G are concrete only if an f-realization is specified. We use $S_A^{G^x}$ to denote the sequence of seed sets selected by S_A^G under an f-realization x.

For an f-realization x and a p-realization y, let Prob[x] (resp. Prob[y]) be the probability with which x (resp. y) happens and Prob[$x|y$] be the probability that x happens conditioned on y.

Definition 7: An f-realization x is *compatible* to a p-realization y if x can be obtained from y by changing the states of some edges in y from {*undetermined*, *selected-undetermined*} into {*selected-live*, *selected-not live*, *not selected*}.

Informally, if x is compatible to y then x is a possible successive state of y in the diffusion process. Similarly, we have the compatibility relationship between two p-realizations. Let ϵ be the empty realization where all the edges are in the *undetermined* state. For a DIC network G, we denote the set of the f-realizations compatible to a p-realization y by $C^G(y)$.

For each concrete strategy $(s_1, \ldots, s_N)$ on $G = (V, E, F_V, F_E)$, there is a corresponding seed set $V' \subseteq \bigcup_{i>0} V_c^i$ in c-G, constructed as follows. If v_i in G is selected by $(s_1, \ldots, s_N)$ for k times, then we add $v_{i,1}, \ldots, v_{i,k}$ in c-G to V'. By this setting, given an f-realization x of c-G, the active nodes under S_A^G in G are those in V_c^0 which are connected to a node in V' via *live* edges in c-G. In the sense of Example. 1, an example f-realization x_1 with strategy $(\{v_3\}, \{v_3\}, \emptyset, \{v_1\})$ is illustrated in Fig. 3.

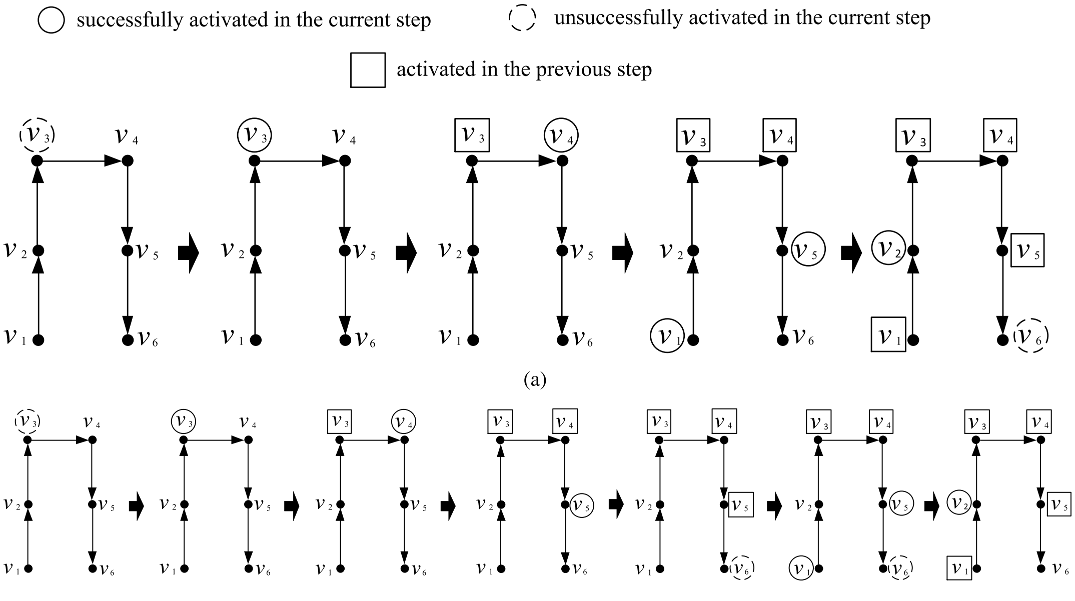

successfully activated in the current step
unsuccessfully activated in the current step
activated in the previous step
v1
v2
v3
v4
v5
v6
(a)
(b)

Fig. 4. Seeding processes of $S_{A'}^{G_1^{x1}}$ and $S_{A^*}^{G_1^{x1}}$. (a) $S_{A'}^{G_1^{x1}}$. (b) $S_{A^*}^{G_1^{x1}}$.

For an f-realization x, let $Node(S_A^{G^x})$ be the union of the corresponding seed sets produced by $S_A^{G^x}$ in c-G in x. For a node-set $V' \subseteq \bigcup_{i>0} V_c^i$, let $N_x^G(V')$ be the number of active nodes in x under seed set V'. Therefore,

$$E[S_A^G] = \sum_{x \in C^G(\epsilon)} \text{Prob}[x] \cdot N_x^G(Node(S_A^{G^x})) \tag{1}$$

$N_x^G(.)$ has the following important properties.

Property 1 (Monotonicity): If $V_1 \subseteq V_2$, then $N_x^G(V_1) \leq N_x^G(V_2)$.

Property 2 (Submodularity): For two node-subsets V_1 and V_2 of $\bigcup_{i>0} V_i$, and a node $v' \in \bigcup_{i>0} V_i$, where $V_1 \subseteq V_2$, $v' \notin V_2$, we have

$$\begin{aligned} &N_x^G(V_2 \cup \{v'\}) - N_x^G(V_2) \\ &\quad \leq N_x^G(V_1 \cup \{v'\}) - N_x^G(V_1). \end{aligned} \tag{2}$$

Proof: Note that the two sides of Eq. (2) represent the number of new active nodes after adding v' to V_2 and V_1, respectively. Suppose u is a new active node incurred by adding v' to V_2, which means u is not connected to any node in V_2 via live edges and u is connected to v' via edges. Since $V_1 \subseteq V_2$, u is not connected to any node in V_1 via live edges either. Therefore, u must also be a new active node when adding v' to V_1. □

B. Optimal Pattern

As introduced in Sec. II-A, a seeding pattern identifies how many budgets should we consume in each step. Now, we show that A^* is the optimal pattern.

Lemma 1: For any DIC network G, suppose A' is an arbitrary seeding pattern and $S_{A'}^G$ is a known seeding strategy of A' on G . There exist a seeding strategy $S_{A^*}^G$ of A^* on G such that $E[S_{A^*}^G] = E[S_{A'}^G]$.

Proof: The main idea is to construct a strategy $S_{A^*}^G$ according to $S_{A'}^G$ such that, in any f-realization x, $N_x^G(Node(S_{A'}^{G^x})) = N_x^G(Node(S_{A^*}^{G^x}))$.

Let $\overline{x}$ be an arbitrary but unknown f-realization of c-G. Suppose $S_{A'}^{G^{\overline{x}}} = (s_1, \ldots, s_N)$ and $A' = (a_1, \ldots, a_N)$. Assume $s_i = \{v_{i,1}, \ldots, v_{i,a_i}\}$ where the nodes are randomly ordered. Note that s_1 is known before the start of diffusion process and s_i ($i > 1$) is unknown until step i as it depends on the outcomes of the past rounds (i.e. the p-realization by the i^{th} step). Let Q be the sequence of the nodes in $\cup s_i$, where the nodes are non-decreasingly ordered by their indexes in s_i according to the lexicographical order. Following pattern A^*, let $S_{A^*}^{G^{\overline{x}}}$ choose the node in Q in order. For the example shown in Fig. 3 with f-realization x_1, the seeding process of strategy $S_{A_1}^{G_1^{x1}}$ and its corresponding strategy $S_{A^*}^{G_1^{x1}}$ are shown in Fig 4.

One can see that $S_{A^*}^{G^{\overline{x}}}$ does nothing but choose the nodes that are chosen by $S_{A'}^{G^{\overline{x}}}$. However, we should node that the same node may not be selected in the same step by $S_{A^*}^{G^{\overline{x}}}$ and $S_{A'}^{G^{\overline{x}}}$. Note that although $S_{A'}^G$ is known to us, the seed nodes produced by $S_{A'}^{G^{\overline{x}}}$ are undetermined as they depends on $\overline{x}$. Suppose $S_{A^*}^{G^{\overline{x}}}$ selects $v_{i,j}$ in the l^{th} step, and the p-realizations under $S_{A'}^{G^{\overline{x}}}$ by step i and that under $S_{A^*}^{G^{\overline{x}}}$ by step l are y_1 and y_2, respectively. To guarantee the feasibility of the construction of $S_{A^*}^{G^{\overline{x}}}$, y_2 must be compatible to y_1. In other words, in realization $\overline{x}$, the events happening by step i under strategy $S_{A'}^{G^{\overline{x}}}$ is a subset of that of happening by step l under strategy $S_{A^*}^{G^{\overline{x}}}$. For otherwise, in step l, $S_{A^*}^G$ cannot determine which node $v_{i,j}$ is.

In fact, such feasibility can be guaranteed by pattern A^*. Let $\overline{v}_i$ be the i^{th} node in Q. Suppose $S_{A'}^{G^{\overline{x}}}$ and $S_{A^*}^{G^{\overline{x}}}$ seeds $\overline{v}_i$ in step l_i' and step l_i^*, respectively. Let y_i' (resp. y_i^*) be the p-realization under $S_{A'}^{G^{\overline{x}}}$ (resp. $S_{A^*}^{G^{\overline{x}}}$) by step l_i' (resp. l_i^*). We need to prove that y_i^* is compatible to y_i', for any $i > 1$.

We prove it by induction. Clearly, y_1^* is compatible to y_1' as $y_1^* = y_1' = \epsilon$. Suppose y_i^* is compatible to y_{l_i}' for any i less than some k. Now we prove that y_k^* is compatible to y_k'. For contraction, suppose y_k^* is not compatible to y_k'. By the supposition, there is an event in x that happens in y_k' while has not happened in $y_{l_k}^*$. However, $y_{l_{k-1}}^*$ is compatible to y_{k-1}', and, by pattern A^*, there is no node can be further activated in realization $\overline{x}$ by step l_k^* under $S_{A^*}^G$. This implies that $S_{A'}^G$ must have waited for some null rounds between step l_{k-1}' and step l_k', which is a contradiction.

By the above construction of $S_{A^*}^G$, since $Node(S_{A^*}^{G^x}) = Node(S_{A'}^{G^x})$ in any f-realization x, we have $E[S_{A^*}^G] = E[S_{A'}^G]$ according to Eq. (1). □

One can see that any strategy of a pattern other than A^* cannot always simulate the one of pattern A^* due to the feasibility issue as discussed above. Intuitively, pattern A^* is the optimal because it maximizes the information obtained before making seeding decision and thus brings us more options in selecting seed nodes. The above result is summarized as follows.

Theorem 1: Pattern A^* is the optimal pattern on any graph G, i.e., for any pattern A, $E[OPT_{A^*}^G] \geq E[OPT_A^G]$.

Proof: By Lemma 1, for any pattern A and network G, we always have some strategy $S_{A^*}^G$ of A^* such that $E[S_{A^*}^G] = E[OPT_A^G]$. Thus,

$$E[OPT_{A^*}^G] \geq E[S_{A^*}^G] = E[OPT_A^G].$$

□

The optimality of pattern A^* is also verified by a simple simulation on a power-law graph with 1000 nodes where the propagation probability of each edge is generated from an exponential distribution with a mean of 0.1. We consider four patterns, A^* and A_i, $1 \leq i \leq 3$, where in pattern A_i we seed i node in each seeding step until the budget is exhausted. The seed nodes are selected according to Rule. 1. For example, for pattern A_2 with a budget of five, we follow the pattern of $(2, 2, 1)$ and in each of the first two seeding steps we select the two nodes that can maximize the active nodes conditioned on the observed influence diffusion.[5] As shown in Fig. 5, the seeding algorithm with A^* performs much better than those with A_1, A_2 and A_3. This is intuitive as A^* holds more information when making seeding decisions. One can also see that A_1 outperforms A_2 and A_3 but A_2 and A_3 are not comparable to each other. Actually, for A_2 and A_3, we cannot determine which one is better without knowing the network structure.

C. Approximation Ratio

In this section, we show that $\overline{S}_{A^*}^G$ has an approximation ratio of $(1 - 1/e)$.

The method for representing the random event space is critical to the analysis of a stochastic model. Essentially, the adaptive seeding strategy $\overline{S}_{A^*}^G$ forms a decision tree, where each node in the tree is a selected seed set and each out-edge of the tree-node represents a possible successive event.

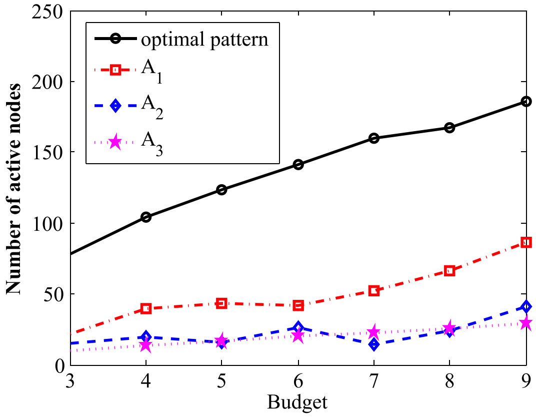


Fig. 5. Comparison between patterns. The y-axis and x-axis denote the number of active nodes and the budget, respectively. The graph gives four curves plotting the influence spreads under four considered patterns with Rule 1, respectively.

Let the root node be the first level. Then, each branch from level i to level $i+1$ corresponds to a p-realization after round i under $\overline{S}_{A^*}^G$. Each path from the root to a leave is a sequence of p-realizations where each p-realization is compatible to its predecessor. For the decision tree of $\overline{S}_{A^*}^G$, let $Z_i = \{z_i^1, \ldots, z_i^{|Z_i|}\}$ be the set of the p-realizations (branches) from level i to level $i+1$ where $|Z_i|$ is number of branches, and $Z_0 = \{\epsilon\}$. Although the basic event space is unique, it can be represented via different decision trees under different strategies. For Example 1 shown in Fig. 1, the decision tree of a strategy of pattern A^* on G_1 is shown in Fig. 6 where the explanations are available in the caption. Note that for a DIC network G the decision tree of $\overline{S}_{A^*}^G$ is determinate.

Now we are ready to show the main result of this paper. Our goal is to prove that

$$E[OPT_{A^*}^G] \leq (1 - 1/e) \cdot E[\overline{S}_{A^*}^G].$$

For an arbitrary network G, let t_i be the i^{th} seed node selected by $OPT_{A^*}^G$, and $T_i = \{t_1, \ldots, t_i\}$. Similarly let w_i be the i^{th} seed node selected by $\overline{S}_{A^*}^G$ and $W_i = \{w_1, \ldots, w_i\}$. Set $T_0 = W_0 = \emptyset$. We use the decision tree to analyze the seeding process.

For a node set V' and a p-realization z_i^j, let

$$F_i^j(V') = \sum_{x \in C_G(z_i^j)} \text{Prob}[x|z_i^j] \cdot N_x^G(V') \tag{3}$$

and

$$F_i(V') = \sum_{j=1}^{|Z_i|} \text{Prob}[z_i^j] \cdot F_i^j(V'). \tag{4}$$

One can see that $F_i^j(W_i)$ is the expected number of active nodes under seed set W_i conditioned on p-realization z_i^j and $F_B(W_B) = E[\overline{S}_{A^*}^G]$.

Therefore, Rule 1 means

$$w_{i+1} = \arg\max_v F_i^j(W_i \cup \{v\}). \tag{5}$$

[5] The reader is referred to Sec. III-D for the detailed implementation of Rule 1.

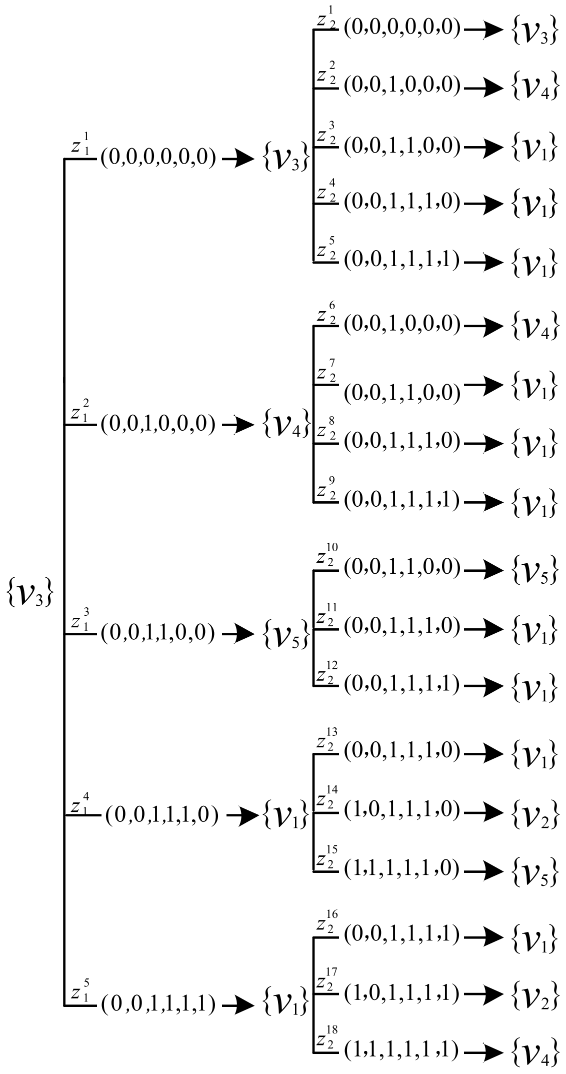

z_2^1 (0,0,0,0,0,0) {v_3}
z_2^2 (0,0,1,0,0,0) {v_4}
z_1^1 (0,0,0,0,0,0) {v_3}
z_2^3 (0,0,1,1,0,0) {v_1}
z_2^4 (0,0,1,1,1,0) {v_1}
z_2^5 (0,0,1,1,1,1) {v_1}
z_2^6 (0,0,1,0,0,0) {v_4}
z_2^7 (0,0,1,1,0,0) {v_1}
z_1^2 (0,0,1,0,0,0) {v_4}
z_2^8 (0,0,1,1,1,0) {v_1}
z_2^9 (0,0,1,1,1,1) {v_1}
z_2^{10} (0,0,1,1,0,0) {v_5}
{v_3}
z_1^3 (0,0,1,1,0,0) {v_5}
z_2^{11} (0,0,1,1,1,0) {v_1}
z_2^{12} (0,0,1,1,1,1) {v_1}
z_2^{13} (0,0,1,1,1,0) {v_1}
z_1^4 (0,0,1,1,1,0) {v_1}
z_2^{14} (1,0,1,1,1,0) {v_2}
z_2^{15} (1,1,1,1,1,0) {v_5}
z_2^{16} (0,0,1,1,1,1) {v_1}
z_1^5 (0,0,1,1,1,1) {v_1}
z_2^{17} (1,0,1,1,1,1) {v_2}
z_2^{18} (1,1,1,1,1,1) {v_4}

Fig. 6. The decision tree of a strategy under pattern A^* on the example DIC network G_1. For the vector $(x_1, x_2, x_3, x_4, x_5)$ on a branch z_i^j, that $x_i = 1$ (resp. $x_i = 0$) means node v_i is active (resp. inactive) after round i through branch z_i^j. In this example, branch z_1^1 implies v_3 is not successfully activated in step 1, and following pattern A^* there are totally 5 and 18 branches from level 1 to level 2 and from level 2 to level 3, respectively.

Let

$$\Delta_i^j = F_i^j(W_i \cup \{w_{i+1}\}) - F_i^j(W_i), \tag{6}$$

for $0 \leq i \ \leq B-1$.

The following lemma shows the difference of $F_i^j(T_B)$ and $F_i^j(W_i)$ can be bound by $B \cdot \Delta_i^j$ due to the submodularity.

Lemma 2:

$$F_i^j(T_B) \leq F_i^j(W_i) + B \cdot \Delta_i^j$$

Proof: See Appendix A-A. □

Let

$$\Delta_i = \sum_{j=1}^{|Z_i|} \text{Prob}[z_i^j] \cdot \Delta_i^j, \tag{7}$$

which is the expected marginal profit at step i. The follow lemma implies the expected active nodes after the i^{th} seeding step is the sum of the marginal profit of the past seeding steps.

Lemma 3: $F_i(W_i) = \Delta_0 + \ldots + \Delta_{i-1}$

Proof: See Appendix A-B. □

Finally, we have the following result.

Theorem 2: $\overline{S}_{A^*}^{G}$ is a strategy within a factor $1 - 1/e$ from the optimal strategy of pattern A^*, i.e., $E[OPT_{A^*}^G] \leq (1 - 1/e) \cdot E[\overline{S}_{A^*}^{G}]$

Proof: For $0 \leq i \ \leq B-1$, we have

$$E[OPT_{A^*}^G] = F_0^1(T_B) = F_i(T_B).$$

By Lemma 2,

$$F_i^j(T_B) \leq F_i^j(W_i) + B \cdot \Delta_i^j,$$

i.e.,

$$E[OPT_{A^*}^G] \leq F_i(W_i) + B \cdot \Delta_i,$$

Thus, combining Lemma 3,

$$E[OPT_{A^*}^G] \leq \Delta_0 + \ldots + \Delta_{i-1} + B \cdot \Delta_i, \tag{8}$$

By multiplying the both sides of Eq. (8) by $(1 - 1/B)^{B-1-i}$ we have

$$\begin{aligned} &E[OPT_{A^*}^G] \cdot (1 - 1/B)^{B-1-i} \\ &\quad \leq (\Delta_0 + \ldots + \Delta_{i-1} + B \cdot \Delta_i) \cdot (1 - 1/B)^{B-1-i} \end{aligned} \tag{9}$$

Now we add up Eq. (9) for $0 \leq i \ \leq B-1$. The left side of the summation is

$$\begin{aligned} &\sum_{i=0}^{B-1} E[OPT_{A^*}^G] \cdot (1 - 1/B)^{B-1-i} \\ &= B(1 - (1 - \frac{1}{B})^B) \cdot E[OPT_{A^*}^G] \end{aligned} \tag{10}$$

On the right side, the coefficient of Δ_i is

$$B \cdot (1 - \frac{1}{B})^{B-i} + \sum_{j=i}^{B-1} (1 - 1/B)^{B-1-j} = B \tag{11}$$

Thus, by Eqs. (10) and (11),

$$\begin{aligned} &E[OPT_{A^*}^G] \cdot B \cdot (1 - (1 - 1/B)^B) \\ &\leq B \cdot (\Delta_0 + \ldots + \Delta_{B-1}) \\ &\quad \{ \text{ by Lemma 3 } \} \\ &= B \cdot F_B(W_B) \\ &= B \cdot E[\overline{S}_{A^*}^{G}]. \end{aligned}$$

Therefore, the approximation ratio of $S_{A^*}^G$ is at least $(1-1/e)$. □

Since A^* is the optimal pattern as discussed in Sec. III-B, $\overline{S}_{A^*}^{G}$ is an $(1 - 1/e)$-approximation of AIM problem.

Corollary 1: $\overline{S}_{A^*}^{G}$ is an $(1 - 1/e)$-approximation of AIM problem.

Golovin and Krause [16] apply the stochastic submodular maximization technique to several applications including the influence diffusion in social networks. They conjecture that applying Rule. 1 to pattern A_0 in the classic IC model yields an $(1 - 1/e)$-approximation to the optimal seeding strategy

Algorithm 1 A-Greedy

```
1: Input: $G = (V, E, F_V, F_E)$ and budget B.
2: CurrentBudget $\leftarrow 0$; $A \leftarrow \emptyset$;
3: $y_0 = \epsilon$;  // $y_i$ is the p-realization after round $i$ .
4: for each $v$ in $V$ do $S_v \leftarrow +\infty$;
5: for $i = 1 : N$ do
6:     if (CurrentBudget<B and no nodes can be further activated) then
7:         for each $v$ in $V \setminus A$ do $s_v \leftarrow false$;
8:         while true do
9:             $v^* = \arg\max_{v \in V \setminus A} S_v$
10:            if ($s_{v^*} = true$) then $A \leftarrow A \cup v^*$; break;
11:            else $s_{v^*} = \sum_{x \in C_G(y_{i-1})}$ Prob$[x|y_{i-1}] \cdot N_x^G(A \cup v^*)$
12:        CurrentBudget=CurrentBudget+1;
13:    Get $y_i$;  // wait for a round of spread
14: $y^* \leftarrow y_N$
15: Return $N_{y^*}^G(A)$
```

under pattern A_0. Actually the derivation of Theorem 2 can be applied to any pattern where we seed at most one node in each step in the DIC model. Therefore, since the classic IC model is a special case of the DIC model, their conjecture in [16] can be confirmed. In fact, under any pattern, Rule 1 is able to provide an approximation with the same ratio. As this paper focuses on practical adaptive seeding strategies, we will not show the technical proof of that result.

D. Implementation Issues

To implement the proposed greedy algorithm, the only problem left is to calculate Eq. (5). Unfortunately, as discussed in [9], it is #P-hard to calculate the real value of $\sum_{x \in C_G(z_j^j)}$ Prob$[x|y] \cdot N_x^G(V')$ in Eq. (3). However, we can employ the Monte Carlo simulation to obtain an accurate estimation. By the Hoeffding's Inequality, the error of the estimation can be infinitely small when a sufficient number of simulations are performed. Thus, with the process of Monte Carlo simulation, the approximation ratio of the above greedy algorithm is $(1 - 1/e + \epsilon)$ where ϵ is the small error incurred by simulation. Another issue one may concern is the efficiency of the greedy algorithm because a large number of simulations may be needed for an accurate estimation. As shown in [22], the Lazy-Forward technique could be implemented in a hill-climbing strategy and leads to far fewer evaluations. The pseudo-code of $\overline{S}_{A^*}^G$ with Lazy-Forward method is shown in Algorithm 1. We denote this adaptive seeding strategy by **A-Greedy**. Note that the time consumed in line 13 is not part of the running time of A-Greedy. In real cases, especially for online social networks, the influence usually spreads very fast [23].

IV. Heuristic Seeding Strategy

In this section, we present a heuristic adaptive seeding strategy based on the greedy algorithm in Sec. III. To reduce the time consumed in the seeding process, a simple idea is to reduce the number of nodes that could be considered as seed nodes. Obviously, the performance of the seeding strategy cannot be guaranteed if we inappropriately exclude some nodes before the seeding process. Thus, it is nature to study that what kinds of nodes can be ignored in the seeding process. An important observation as shown later in Sec. V is that there could be a significant gap between the strength of influential nodes and other nodes. This fact is coincident to the power-law nature of the real-world social networks where degree of the nodes follows the exponential distribution. Motivated by this observation, we design a heuristic seeding strategy, termed as **H-Greedy**, which narrows the candidate seed set before the seeding process.

H-Greedy: Let $H(v)$ be the number of the nodes can be activated by a single seed node v. Let E[.] and Std[.] denote the mean and the standard deviation of a random variable. H-Greedy consists of two steps. First, before we start the seeding process, by Monti Carlo simulation, we first obtain the estimates of $E[H(v)]$, $E[\sum_{v \in V} H(v)/N]$, and $Std[\sum_{v \in V} H(v)/N]$. We denote those three estimates by $\hat{\text{E}}[H(v)]$, $\hat{\text{E}}[\sum_{v \in V} H(v)/N]$, and $\hat{\text{S}}\text{td}[\sum_{v \in V} H(v)/N]$, respectively. Then, we run A-Greedy and omit the node v if $E[H(v)]$ is less than the lower 1-sigma control[6] of $\sum_{v \in V} H(v)/N$.

As discussed in the prior works, we used to execute Monte Carlo simulation for 10000 to 20000 times for an accurate estimation. However, in the first step of H-Greedy, 1000 to 2000 simulations are sufficient. This is because the estimates are not necessary to be very accurate as they are merely used to narrow the candidate set of seed nodes. With a smaller set of candidate seed nodes the time consumed in the seeding process can be significantly reduced. As shown later, the performance of H-Greedy is close to that of A-Greedy which has a provable performance guarantee. We will further discuss the feasibility of H-Greedy in the next section.

V. Experiment

In this section, we show the results of the conducted experiments. We evaluate the proposed seeding strategies from the following aspects: (a) the influence spread of A-Greedy compared to that of non-adaptive seeding strategies; (b) the effectiveness and efficiency of the heuristic strategy.

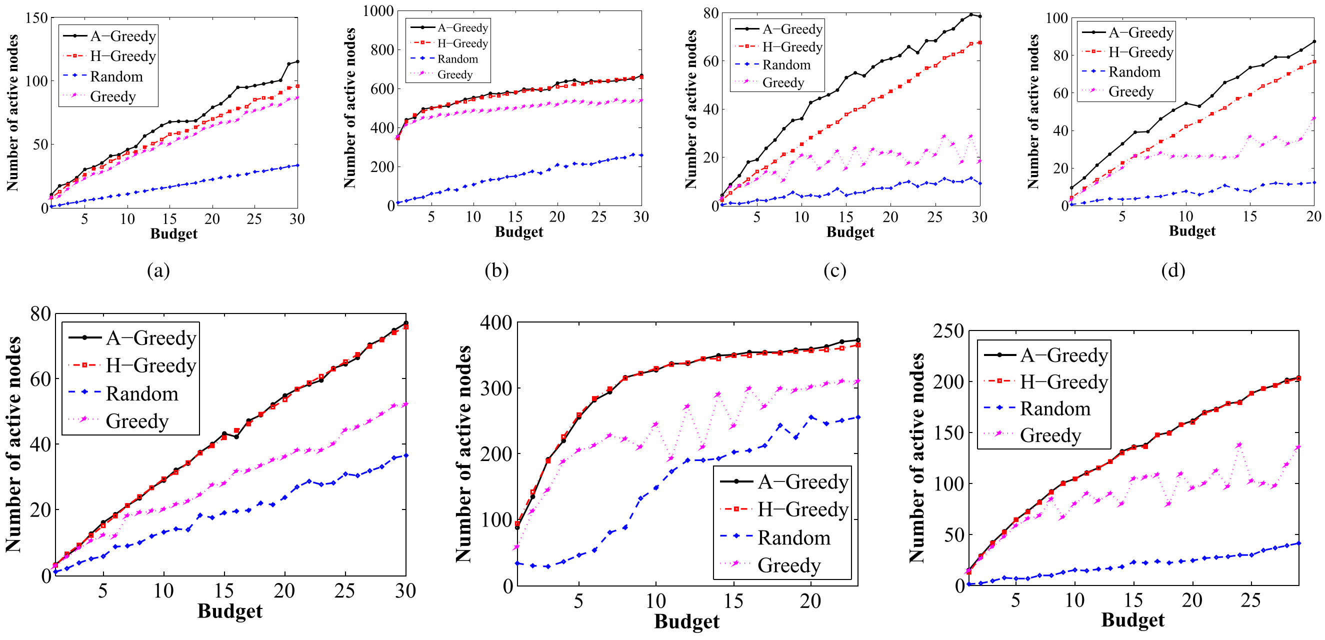
A. Setup

In order to fairly compare our seeding strategies to the existing approaches, we employ two real-world social networks, which have been widely used in the prior works, and a synthetic power-law network which is able to capture the key features of real social networks. The propagation probabilities are generated from three distributions, as shown later.

Network Structure: The first real-world social network, denoted by **Hep**, is an academic collaboration from co-authorships in physics. Hep is compiled from the "High Energy Physics - Theory" section of the e-print arXiv[7] and has

[6] Mean minus standard deviation.

[7] http://www.arXiv.org



(a) (b) (c) (d)

(e) (f) (g)

Fig. 7. Comparing A-Greedy with Greedy. In all seven graphs, the y-axis and x-axis denote the number of active nodes and the budget, respectively. Each graph gives four curves plotting the influence spread under four seeding strategies, respectively. (a) $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$. (b) $\mathscr{F}^3$ with Prob$[X_u = 1] = 1$ (c) $\mathscr{F}^1$ with Prob$[X_u = 1] = 0.5$ (d) $\mathscr{F}^2$ with Prob$[X_u = 1] = 0.5$ on Hep. (e) $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$ on PL. (f) $\mathscr{F}^3$ with Prob$[X_u = 1] = 0.5$ on PL. (g) $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$ on Wiki.

been widely used in the prior works (e.g. [5], [9], [11], [24]). For each pair of authors who has a co-authorship, there are two directed edges from each one to the other. The resulted network has about 15,000 nodes and 58,000 directed edges. The second dataset, denoted by **Wiki**, contains the Wikipedia voting data [25] from the inception of Wikipedia. Nodes in this network represent wikipedia users and a directed edge from node u to node v represents that user u votes on user v, which mean v has influence over u. Thus, if there is an edge from u to v in the original data, we add an edge from v to u in Wiki. Wiki has about 8,600 nodes and 103,000 directed edges and has been studied in [26]–[28]. The last dataset is a synthetic power-law network generated by DIGG [29]. The synthetic power-law network selected in this paper, denoted by **PL**, includes 2500 nodes and 26,000 directed edges. Power-law degree distribution has been shown to be one of the most important characteristics of social networks [21]. We use PL dataset to evaluate the performance of the proposed seeding strategies in general social networks.

Propagation Probability: We uniformly assign a distribution of the propagation probability for each edge. The three distributions considered are shown as follows. In $\mathscr{F}^1$, the propagation probability are fixed as 0.01, which is the same as that in [5]. $\mathscr{F}^2$ is an exponential distributions with a mean of 0.01. $\mathscr{F}^3$ is a uniform discrete distribution over $\{0.1, 0, 01, 0, 001\}$.

Activation Probability: We uniformly assign an activation probability on each node u, and set Prob$[X_u = 1]$ to be 1 and 0.5.

Note that it reduces to the classic IC model under $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$.

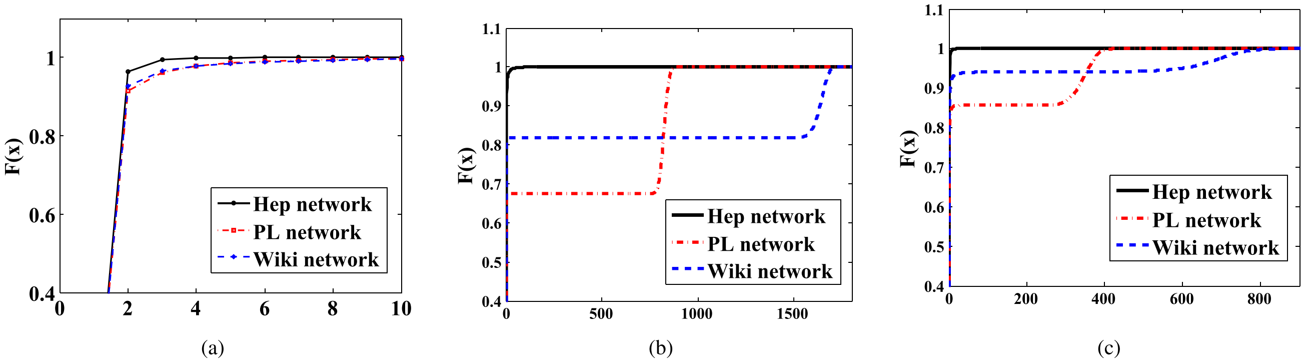
Seeding Strategies: The tested seeding strategies are shown as follows.

1) **Greedy**. This is the state-of-art non-adaptive seeding approach proposed in [5]. In Greedy, the nodes are selected by a hill-climbing algorithm before the diffusion process. When implementing Greedy in the DIC model, we fix the propagation probability by its mean as the real propagation probabilities are unavailable in the DIC model before the start of diffusion process. For each estimation, 10000 simulations are run to obtain an accurate estimate.
2) **A-Greedy**. This is the greedy adaptive seeding strategy proposed in Sec. III. Similarly, 10000 simulations are run to obtain an accurate estimate of $\sum_{x \in C_G(y_{i-1})}$ Prob$[x|y_{i-1}] \cdot N_x^G(A \cup v^*)$ in line 11 of Algorithm 1.
3) **H-Greedy**. This is the heuristic adaptive seeding strategy proposed in Sec. IV. In the first step of H-Greedy, 2000 simulations are run to obtain the estimates mentioned in Sec. IV.
4) **Random**. This is a baseline seeding strategy where the seed nodes are selected randomly.

As discussed in the prior works, the seeding strategies based on the shortest-path and high-degree perform worst than Greedy. Thus we ignore other seeding strategies. In our experiment, the budget is chosen from 10 to 30.

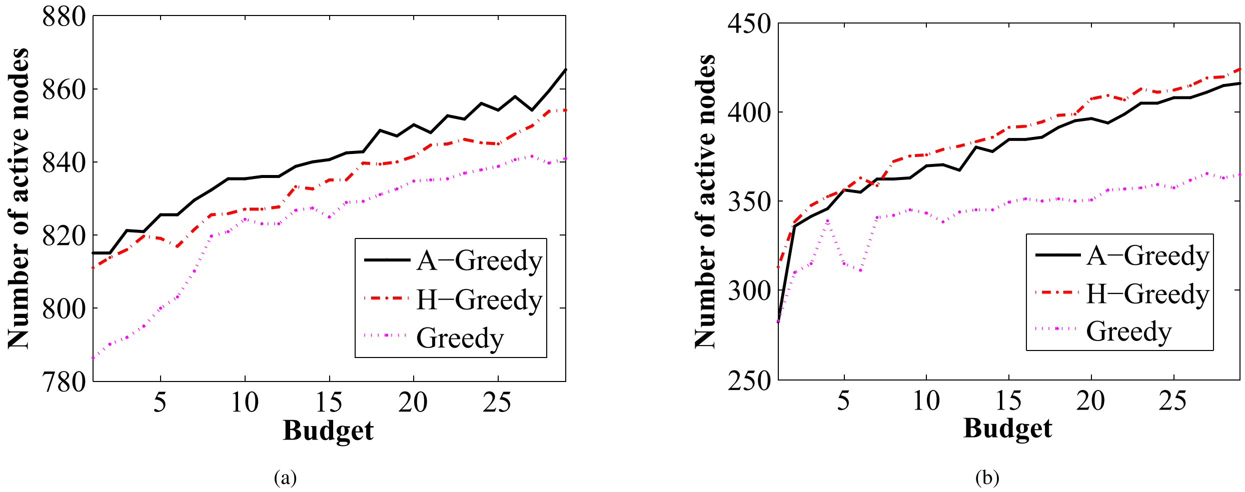
B. Results

First, we discuss the performance of A-Greedy. As shown in Fig. 7, A-Greedy outperforms Greedy under all circumstances.



(a) (b) (c)

Fig. 8. Distributions of $E(H(v))$ in the three networks under different propagation probability. (a) $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$. (b) $\mathscr{F}^2$ with Prob$[X_u = 1] = 1$. (c) $\mathscr{F}^3$ with Prob$[X_u = 1] = 1$.



(a) (b)

Fig. 9. Comparing H-Greedy with A-Greedy. The y-axis and x-axis denote the number of active nodes and the budget, respectively. Each graph gives three curves plotting the influence spreads under A-Greedy, H-Greedy and Greedy, respectively. We ignore Random here as it performs poorly. (a) $\mathscr{F}^2$ with Prob$[X_u = 1] = 1$ on PL. (b) $\mathscr{F}^3$ with Prob$[X_u = 1] = 1$ on PL.

This is intuitive as the adaptive seeding strategies are able to utilize the outcomes of the past rounds. As shown in Fig. 7a, A-Greedy is superior to Greedy by a notable margin even in the classic IC model. For the DIC model where the diffusion process is of more uncertainness, the results herein verify the significant advantages of the adaptive seeding strategy over the non-adaptive seeding strategy. We discuss the results in detail in the following.

For the Hep network, as shown in Fig. 7a, A-Greedy is 125% better than Greedy in the classic IC model under $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$. While the uncertainness of the diffusion process getting increased, namely by changing Prob$[X_u = 1]$ to 0.5 as shown in Fig. 7c, A-Greedy becomes 320% better than Greedy. As shown in Figs. 7e, 7f and 7g, for PL and Wiki network, we have the similar result. For example, for the PL network under $\mathscr{F}^1$ with Prob$[X_u = 1] = 1$, as shown in Fig. 7e, one seed node results about 2.5 active nodes under A-Greedy while in average 1.67 nodes can be activated by a single seed node under Greedy. Another important observation is that the curves generated by Greedy become less stable in the DIC model, which implies that to reach the same level of accuracy Greedy requires more simulations than A-Greedy does.

Now let us discuss the performance of the proposed heuristic seeding strategy H-Greedy. Fig. 8 shows the distribution of $E[H(v)]$ drew from the datasets by simulation. In Fig. 8a, 90 % of the nodes cannot activate more than 2 nodes, while in Figs. 8b and 8c, we can see that there is a significant gap between the strength of influential nodes and other nodes. For example, as shown in Fig. 8b, 24 percent of the nodes in Wiki could activate more than 1600 nodes while 82 percent of them can hardly activate more than 50 nodes. For the PL network in the same setting, about 30 percent of the nodes could bring 780 active nodes while 68 percent of them only result less than 100 active nodes. Admitting that the difference of $E[H(v)]$ between two nodes would decrease along with the seeding process due to the submodularity feature, the nodes with small $E[H(v)]$ are not likely to be selected as a seed node as the gap is too large and we only have a small budget compared to the population of users. Thus, 1-sigma control on $E[H(v)]$ is a safe bound such that we will not miss any influential nodes. For the circumstances in Fig. 7 the performance of

TABLE IV

SCALABILITY OF H-GREEDY. THE FOUR CASES ARE SHOWN IN THE FIRST COLUMN. THE SECOND AND THIRD COLUMNS SHOW THE AVERAGE TIME CONSUMED IN SELECTING ONE SEED NODE UNDER H-GREEDY AND A-GREEDY, RESPECTIVELY

Parameter Setting	**H-Greedy** (ms)	**A-Greedy** (ms)
$\mathscr{F}^2$ & Prob$[X_u = 1] = 1$ on PL	14977	51485
$\mathscr{F}^2$ & Prob$[X_u = 1] = 1$ on Wiki	87412	268499
$\mathscr{F}^3$ & Prob$[X_u = 1] = 1$ on PL	981	11931
$\mathscr{F}^3$ & Prob$[X_u = 1] = 1$ on Wiki	31247	44625

H-Greedy is almost the same as that of A-Greedy. This is because in those settings H-Greedy can hardly eliminate any node as the patterns of the distributions of $E[H(v)]$ are like Fig. 8a. However, when the distribution of $E[H(v)]$ has a pattern like that in Figs 8b or 8c, H-Greedy would be an effective and efficient strategy. In these cases, H-Greedy could rule out more than a half of the nodes and thus 20% to 90% of the time consumed in the seeding process could be reduced as shown in Table IV. Furthermore, as shown in Figs 9a and 9b, H-Greedy performs slightly worse than A-Greedy but still better than Greedy. Note that the feature of the distribution of $H(v)$ shown in Figs 8b and 8c is critical to H-Greedy. It is interesting to further design heuristic strategies with more sophisticated pruning methods.

VI. CONCLUSION AND FUTURE WORK

In this paper we have considered the problem that how to maximize the spread of influence in dynamic social networks. The proposed DIC model is able to capture the dynamic aspects of a real social network and the uncertainness of the diffusion process. In the DIC model, a certain node can be seeded for more than once and the propagation probability between two users varies following a certain distribution. Based on the DIC model, we formulate the adaptive seeding strategies by introducing the concept of seeding pattern. The pattern A^* constructed in Sec. II shows the optimal method of determining how many budgets shall we utilize in each seeding step. Combining the optimal pattern with the natural hill-climbing algorithm, we present the A-Greedy seeding strategy and show that A-Greedy has a performance ratio of $(1 - 1/e)$. By the observation that the influential nodes are much more powerful than other nodes in a social network, we further design a simple heuristic adaptive seeding strategy H-Greedy based on A-Greedy. The experimental results herein demonstrate the superiority of the adaptive seeding strategies to prior approaches.

The future work following this topic consists of several aspects. As we can see, H-Greedy is a simple heuristic strategy but not effective for all the settings of DIC model. Thus, we plan to design better heuristic adaptive seeding strategies that are able to deal with general social networks. We note that the technique in [8] is possibly applicable to the adaptive seeding framework and we leave this part as future work. Another aspect of the future work is to design adaptive seeding strategies which are able to meet the round limit. In real applications, we may only concern the spread influence within a certain number of rounds. In this case, the analysis of the adaptive seeding strategies becomes intricate. On the one hand as shown by pattern A^* we try to utilize the budgets as late as possible in order to obtain more information before making seeding decision, while on the other hand, since there is a round limit, delaying a seeding step leads the loss of a diffusion round. Therefore, the optimal pattern proposed in Sec. III-B may not be a good choice under this setting as it tends to result more rounds of diffusion. One can easily check that with a round limit our objective function is not submodular anymore, which renders it difficult to find a greedy algorithm with a provable performance guarantee.

APPENDIX A PROOFS

A. Proof of Lemma 2

Proof: For $1 \leq h \ \leq B$, by Property 2,

$$N_x^G(T_h \cup W_i) - N_x^G(T_{h-1} \cup W_i) \\ \leq N_x^G(\{t_h\} \cup W_i) - N_x^G(W_i)$$

Thus,

$$\begin{aligned} &\sum_{x \in C_G(z_i^j)} \text{Prob}[x|z_i^j]\big(N_x^G(T_h \cup W_i) - N_x^G(T_{h-1} \cup W_i)\big) \\ &\quad \leq \sum_{x \in C_G(z_i^j)} \text{Prob}[x|z_i^j]\big(N_x^G(\{t_h\} \cup W_i) - N_x^G(W_i)\big) \\ &\qquad \{\text{by Eq. (3)}\} \\ &\quad = F_i^j(\{t_h\} \cup W_i)) - F_i^j(W_i) \\ &\qquad \{\text{by Eq. (5)}\} \\ &\quad \leq F_i^j(W_{i+1})) - F_i^j(W_i) \\ &\quad = \Delta_i^j. \end{aligned}$$

Adding the above inequalities for all $1 \leq h \ \leq B$, we have

$$\begin{aligned} &\sum_{1 \leq h \ \leq B} \sum_{x \in C_G(z_i^j)} \text{Prob}[x|z_i^j]\big(N_x^G(T_h \cup W_i) - N_x^G(T_{h-1} \cup W_i)\big) \\ &\quad = \sum_{x \in C_G(z_i^j)} \text{Prob}[x|z_i^j]\big(N_x^G(T_B \cup W_i) - N_x^G(T_0 \cup W_i)\big) \\ &\quad = F_i^j(T_B \cup W_i) - F_i^j(W_i) \\ &\quad \leq B \ \cdot \Delta_i^j. \end{aligned}$$

Thus,

$$F_i^j(T_B) \leq F_i^j(T_B \cup W_i) \leq F_i^j(W_i) + B \ \cdot \Delta_i.$$

Note that T_B depends on x and W_i depends on z_i^j. □

B. Proof of Lemma 3

Proof: Note that, for any $0 \leq h < B$

$$\begin{aligned} F_h(W_h) &= \sum_{j=1}^{|z_h|} \text{Prob} \cdot [z_h^j] F_i^j(W_h) \\ &= \sum_{j=1}^{|z_{h-1}|} \text{Prob}[z_{h-1}^j] \cdot F_{h-1}^j(W_{h-1} \cup \{w_h\}). \end{aligned}$$

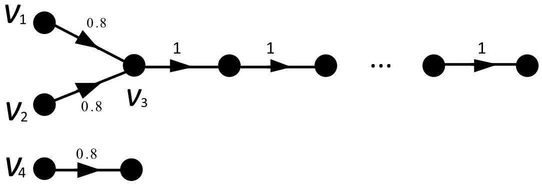


Fig. 10. Example network.

Thus, we have

$$\begin{aligned}
&\Delta_0 + \ldots + \Delta_{i-1} \\
&\quad \{\text{by Eq. (7)}\} \\
&\quad = \sum_{h<i} \sum_{j=1}^{|Z_h|} \text{Prob}[z_h^j] \cdot \Delta_h^j \\
&\quad = \sum_{h<i} \sum_{j=1}^{|Z_h|} \text{Prob}[z_h^j] \cdot (F_h^j(W_h \cup \{w_{h+1}\}) - F_h^j(W_h)) \\
&\quad = \sum_{h<i} (F_{h+1}(W_{h+1}) - F_h(W_h)) \\
&\quad = F_i(W_i) - \sum_{j=1}^{|z_0|} \text{Prob}[z_0^j] \cdot F_0^j(W_0) \\
&\quad = F_i(W_i).
\end{aligned}$$

□

Appendix B

In this appendix, we provide an example to show how the failure of seeding may affect an adaptive seeding algorithm. Consider the example shown in Fig. 10 where v_1, v_2 and v_4 can be seeded with a probability of 0.8 and other nodes cannot be activated as seed nodes. As shown in the figure, v_3 is an influential node which is able to activate many users. Suppose we follow a seeding pattern of (2,1) where two nodes are seeded firstly and then another node is seeded after one round of spread. For the network shown in Fig. 10, we will select v_1 and v_2 in the first seeding step as we hope that v_3 can be later activated. If neither of v_1 and v_2 is activated in the first round, then we should choose one of them as the seed node for the second seeding step; if one or more of v_1 and v_2 are activated, then we will select v_4 in the second step. One can see that whether a node will be repeatedly seeded depends on the network structure and the results of past diffusion rounds.

References

[1] V. Mahajan, E. Muller, and F. M. Bass, "New product diffusion models in marketing: A review and directions for research," *J. Marketing*, vol. 54, no. 1, pp. 1–26, 1990.

[2] J. Goldenberg, B. Libai, and E. Muller, "Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata," *Acad. Marketing Sci. Rev.*, vol. 2001, no. 9, pp. 1–19, 2001.

[3] R. M. Bond *et al.*, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, no. 7415, pp. 295–298, 2012.

[4] L. Fan *et al.*, "Least cost rumor blocking in social networks," in *Proc. IEEE 33rd Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Jul. 2013, pp. 540–549.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. 9th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2003, pp. 137–146.

[6] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 57–66.

[7] J. Goldenberg, B. Libai, and E. Muller, "Talk of the network: A complex systems look at the underlying process of word-of-mouth," *Marketing Lett.*, vol. 12, no. 3, pp. 211–223, 2001.

[8] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 1029–1038.

[9] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2009, pp. 199–208.

[10] N. Chen, "On the approximability of influence in social networks," *SIAM J. Discrete Math.*, vol. 23, no. 3, pp. 1400–1415, 2009.

[11] C. Long and R. C.-W. Wong, "Minimizing seed set for viral marketing," in *Proc. IEEE 11th Int. Conf. Data Mining (ICDM)*, Dec. 2011, pp. 427–436.

[12] N. Du, L. Song, M. Gomez-Rodriguez, and H. Zha, "Scalable influence estimation in continuous-time diffusion networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3147–3155.

[13] M. Gomez-Rodriguez and B. Schölkopf. (2012). "Influence maximization in continuous time diffusion networks." [Online]. Available: http://arxiv.org/abs/1205.1682

[14] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Math. Program.*, vol. 14, no. 1, pp. 265–294, 1978.

[15] A. Asadpour, H. Nazerzadeh, and A. Saberi, "Stochastic submodular maximization," in *Proc. 4th Int. Workshop WINE*, Shanghai, China, Dec. 2008, pp. 477–489.

[16] D. Golovin and A. Krause. (2010). "Adaptive submodularity: Theory and applications in active learning and stochastic optimization." [Online]. Available: http://arxiv.org/abs/1003.3967

[17] L. Seeman and Y. Singer, "Adaptive seeding in social networks," in *Proc. IEEE 54th Annu. Symp. Found. Comput. Sci. (FOCS)*, Oct. 2013, pp. 459–468.

[18] T. Horel and Y. Singer, "Scalable methods for adaptively seeding a social network," in *Proc. 24th Int. Conf. World Wide Web*, 2015, pp. 441–451.

[19] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining*, 2010, pp. 241–250.

[20] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. 12th Int. Conf. KES III*, Zagreb, Croatia, Sep. 2008, pp. 67–75.

[21] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, 2009.

[22] J. Leskovec *et al.*, "Cost-effective outbreak detection in networks," in *Proc. 13th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2007, pp. 420–429.

[23] J. Yang and S. Counts, "Predicting the speed, scale, and range of information diffusion in Twitter," in *Proc. ICWSM*, vol. 10. 2010, pp. 355–358.

[24] P. Zhang, W. Chen, X. Sun, Y. Wang, and J. Zhang, "Minimizing seed set selection with probabilistic coverage guarantee in a social network," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 1306–1315.

[25] J. Leskovec. *Wikipedia Vote Network*. [Online]. Available: http://snap.stanford.edu/data/wiki-Vote.html

[26] W. Chen *et al.*, "Influence maximization in social networks when negative opinions may emerge and propagate," in *Proc. SDM*, vol. 11. 2011, pp. 379–390.

[27] S. Li *et al.*, "Influence maximization in social networks with user attitude modification," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2014, pp. 3913–3918.

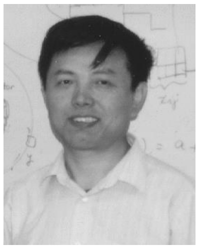
[28] S. Li, Y. Zhu, D. Li, D. Kim, and H. Huang, "Rumor restriction in online social networks," in *Proc. IEEE 32nd IPCCC*, Dec. 2013, pp. 1–10.

[29] L. Cowen, A. Brady, and P. Schmid. *DIGG: Dynamic Graph Generator*. [Online]. Available: http://digg.cs.tufts.edu/

Guangmo Tong received the B.S. degree in mathematics and applied mathematics from the Beijing Institute of Technology in 2013. He is currently pursuing the Ph.D. degree with the Department of Computer Science, The University of Texas at Dallas. His research interests include social networks, data communication, and real-time systems. He is a Student Member of the IEEE.

Weili Wu (M'00) received the M.S. and Ph.D. degrees from the Department of Computer Science, University of Minnesota, Minneapolis, MN, USA, in 1998 and 2002, respectively. She is currently a Full Professor with the Department of Computer Science, The University of Texas at Dallas, Dallas, TX, USA. Her research mainly deals in the general research area of data communication and data management. Her research focuses on the design and analysis of algorithms for optimization problems that occur in wireless networking environments and various database systems.

Shaojie Tang received the Ph.D. degree in computer science from the Illinois Institute of Technology in 2012. He is currently an Assistant Professor with the Naveen Jindal School of Management, The University of Texas at Dallas. His research interest includes social networks, mobile commerce, game theory, e-business, and optimization. He received the Best Paper Awards in ACM MobiHoc 2014 and the IEEE MASS 2013. He also received the ACM SIGMobile Service Award in 2014. He served in various positions (as the Chair and TPC Member) at numerous conferences, including ACM MobiHoc and the IEEE ICNP. He is an Editor of *Information Processing in Agriculture* and the *International Journal of Distributed Sensor Networks*.



Ding-Zhu Du received the M.S. degree from the Chinese Academy of Sciences in 1982, and the Ph.D. degree from the University of California at Santa Barbara in 1985, under the supervision of Prof. R. V. Book. Before settling at The University of Texas at Dallas, he was a Professor with the Department of Computer Science and Engineering, University of Minnesota. He was with the Mathematical Sciences Research Institute, Berkeley, for one year; the Department of Mathematics, Massachusetts Institute of Technology, for one year; and the Department of Computer Science, Princeton University, for one and a half years. He graduated 40 Ph.D. students under his supervision. He is a member of the IEEE. He is the Editor-in-Chief of the *Journal of Combinatorial Optimization* and is also on the Editorial Boards for several other journals.