Influential Nodes in a Diffusion Model for Social Networks

David Kempe[1,⋆], Jon Kleinberg[2,⋆⋆], and Éva Tardos[2,⋆⋆⋆]

[1] Department of Computer Science,
University of Southern California
dkempe@usc.edu
[2] Department of Computer Science, Cornell University
kleinber@cs.cornell.edu, eva@cs.cornell.edu

Abstract. We study the problem of maximizing the expected spread of an innovation or behavior within a social network, in the presence of "word-of-mouth" referral. Our work builds on the observation that individuals' decisions to purchase a product or adopt an innovation are strongly influenced by recommendations from their friends and acquaintances. Understanding and leveraging this influence may thus lead to a much larger spread of the innovation than the traditional view of marketing to individuals in isolation.

In this paper, we define a natural and general model of influence propagation that we term the *decreasing cascade model*, generalizing models used in the sociology and economics communities. In this model, as in related ones, a behavior spreads in a cascading fashion according to a probabilistic rule, beginning with a set of initially "active" nodes. We study the *target set selection* problem: we wish to choose a set of individuals to target for initial activation, such that the cascade beginning with this active set is as large as possible in expectation. We show that in the decreasing cascade model, a natural greedy algorithm is a $1-1/\mathrm{e}-\varepsilon$ approximation for selecting a target set of size k.

1 Introduction

Suppose that we are trying to market a product, or promote an idea, innovation or behavior, within a population of individuals. In order to do so, we can "target" individuals; for instance, this "targeting" could take the form of offering free samples of the product, demonstrating an innovation, or explaining an idea (such as the consequences of drug use to teenagers). An important question is

⋆ This research was supported by an Intel Graduate Fellowship and an NSF Graduate Research Fellowship.
⋆⋆ Supported in part by a David and Lucile Packard Foundation Fellowship and NSF grants 0311333 and 0329064.
⋆⋆⋆ Supported in part by NSF ITR grant CCR-0325453, NSF grant CCR-0311333, and ONR grant N00014-98-1-0589.

L. Caires et al. (Eds.): ICALP 2005, LNCS 3580, pp. 1127–1138, 2005.

then whom we should target. Clearly, if there were no interaction between the individuals, this would be straightforward: the effect on each targeted individual could be determined in isolation, and we could choose the set of individuals with largest (expected) revenue or reach. However, individuals do not exist in a vacuum; rather, they form complex social networks based on a multitude of different relations and interactions. By virtue of these interactions, they influence each other's decisions in adopting a product or behavior.

Research in the area of *viral marketing* [1, 2, 3, 4, 5] takes advantage of these social network effects, based on the premise that targeting a few key individuals may lead to strong "word-of-mouth" effects, wherein friends recommend a product to their friends, who in turn recommend it to others, and so forth, creating a cascade of recommendations. In this way, decisions can spread through the network from a small set of initial adopters to a potentially much larger group. Given a probabilistic model for the way in which individuals influence one another, the *influence maximization problem* consists in determining a set A of k individuals yielding the largest expected cascade.

The influence maximization problem has been proposed and studied by Domingos and Richardson [2, 5], who gave heuristics for the problem in a very general descriptive model of influence propagation. In recent work [6], we obtained provable performance guarantees for approximation algorithms in several simple, concrete, but extensively studied models from mathematical sociology (see, e.g., [7, 8, 9] for comprehensive introductions to this area).

In this paper, we show that the influence maximization problem can be approximated in a very general model that we term the *decreasing cascade model*. The analysis techniques from our earlier work [6] rely on the concrete forms of influence used in that paper, and we show that they cannot be applied to the general model considered here. We therefore develop a more general framework, which we believe will be of interest in its own right, for reasoning about dynamic processes in network models such as these.

1.1 The Decreasing Cascade Model

Throughout this paper, we call individuals (nodes) *active* if they have adopted the product, and *inactive* otherwise. We assume that once a node becomes active, it will remain so forever (see [6] for a discussion on how this assumption can be lifted). We focus on *cascade models* that capture the dynamics of recommendations in a step-by-step fashion: when a node u first becomes active, say at time t, it is considered *contagious*. It has one chance of influencing each previously inactive neighbor v. A successful attempt will cause v to become active in the next time step $t + 1$. If multiple neighbors of v become active at time t, then their activation attempts are sequenced in an arbitrary order, but we assume that they all happen within time step t. After a node u has made all its attempts at influencing other nodes, it remains active, but is now *non-contagious*. The process terminates when there are no more contagious nodes.

In order to fully describe the model, we need to specify the probability of success for node u's attempt at activating v. In the simplest *independent cascade*

model [3], this probability is a constant $p_v(u)$, independent of the history of the process. In general, however, v's propensity for being activated may change as a function of which of its neighbors have already attempted (and failed) to influence it; if S denotes the set of v's neighbors that have already attempted to influence v, then u's *success probability* is denoted by $p_v(u, S)$. For this model to be well-defined, we also need to assume *order-independence*: if all nodes from a set T try to influence v, then the order in which their attempts are made does not affect the probability of v being active in the end. Formally, if $u_1, \ldots, u_r$, and $u'_1, \ldots, u'_r$ are two permutations of T, and $T_i = \{u_1, \ldots, u_{i-1}\}$ as well as $T'_i = \{u'_1, \ldots, u'_{i-1}\}$, then order-independence means that

$$\prod_{i=1}^{r}(1 - p_v(u_i, S \cup T_i)) = \prod_{i=1}^{r}(1 - p_v(u'_i, S \cup T'_i))$$

for all sets S disjoint from T.

From the point of view of influence maximization, we start by *targeting* a set A of individuals for activation at time 1, making them contagious. Afterwards, the process unfolds as described above, until there are no more contagious nodes; we say that the process *quiesces*. Note that this happens after at most $n + 1$ rounds. At that point, we have some set $\varphi(A)$ of active nodes, which is is a random variable. The goal is to choose A so as to maximize the expected size $\sigma(A) := \mathrm{E}[|\varphi(A)|]$ of this final set of active nodes. Due to the computational difficulty of this goal (see the discussion below), we will consider approximation algorithms: for a constant c, we wish to choose a set A for which $\sigma(A)$ is at least $\frac{1}{c}$ times as large as $\sigma(A^*)$ for *any* set A^* of k nodes. The quantity c is thus the approximation guarantee of the algorithm.

The *order-independent cascade model* is very general — it specifies how each node influences each other node, and how the influence is "attenuated" by previous interactions a node has had. It is also equivalent in a precise sense to a generalization of Granovetter's threshold model [10] for social networks (see Section 3).

In general, it is NP-hard to approximately maximize the size $\sigma(A)$ of the final active set to within $n^{1-\varepsilon}$, for any $\varepsilon > 0$. The inapproximability follows from a straightforward reduction, e.g., from VertexCover, and can already be shown in the case of a *hard threshold* model [11, 12, 13], where a node v is activated if at least a fixed fraction (say, 1/2) of its neighbors are active; this corresponds to $p_v(u, S)$ being 0 if S contains fewer than half of v's neighbors, and 1 otherwise.

Thus, we study here a natural restriction that we term the *decreasing cascade model*. In the decreasing cascade model, the functions $p_v(u, S)$ are non-increasing in S, i.e., $p_v(u, S) \geq p_v(u, T)$ whenever $S \subseteq T$. Intuitively, this restriction states that a contagious node's probability of activating some $v \in V$ decreases if more nodes have already attempted to activate v, and v is hence more "marketing-saturated". The decreasing cascade model contains the *independent cascade model* [3] as a special case, and even for the independent cascade model, maximizing $\sigma(A)$ is NP-hard [6]; in fact, the proof in [6] shows that it is NP-hard to approximate within $1 - 1/\mathrm{e} + \varepsilon$ for any $\varepsilon > 0$.

2 An Approximation Algorithm

In this paper, we analyze the following simple greedy algorithm (Algorithm 1.) for influence maximization. The approximation guarantee for this algorithm is the main theorem of this paper:

Algorithm 1. Greedy Approximation Algorithm

1: Start with $A = \emptyset$
2: **for** $i = 1$ to k **do**
3: Let v_i be a node (approximately) maximizing the *marginal gain* $\sigma(A \cup \{v\}) - \sigma(A)$.
4: Set $A \leftarrow A \cup \{v_i\}$.
5: **end for**

Theorem 1. *Let A^* be the the set maximizing $\sigma(\cdot)$ among all sets of k nodes.*

1. *If the optimal v_i is chosen in each iteration, then the greedy algorithm is a $(1 - 1/\mathrm{e})$-approximation, i.e., the set A found by the algorithm satisfies $\sigma(A) \geq (1 - 1/\mathrm{e}) \cdot \sigma(A^*)$.*
2. *If the node v_i is a $1 - \varepsilon$ approximate best node in each iteration, then the greedy algorithm is a $(1 - 1/\mathrm{e} - \varepsilon')$-approximation, where ε' depends on ε polynomially.*

Before proceeding with the proof of Theorem 1, a few words are in order about determining the node v_i in the **for** loop of the algorithm. Even in the simple independent cascade model, it is not clear how to evaluate $\sigma(A)$ exactly, or whether this can be done in polynomial time; in fact, we consider the question of evaluating $\sigma(A)$ an interesting direction for further research. However, the cascade process has the property that it can be efficiently simulated, simply by running the probabilistic rule for influence propagation until quiescence (which, as noted above, will occur within at most $n + 1$ rounds). By repeatedly simulating the cascade process and sampling $\varphi(A)$, we can compute arbitrarily close approximations to $\sigma(A)$. A straightforward calculation shows that with a number of simulations polynomial in ε, δ, and n, one can obtain a $1 \pm \varepsilon$ approximation to $\sigma(A)$, with probability at least $1 - \delta$. This approximate evaluation of $\sigma(A)$ in turn is enough to find an element v whose marginal gain $\sigma(A \cup \{v\}) - \sigma(A)$ is within a factor of $1 - \varepsilon'$ of maximal.

The idea for the proof of Theorem 1 is to show that $\sigma(A)$ is a monotone and submodular function of A. The property of submodularity formally means that $\sigma(S \cup \{w\}) - \sigma(S) \geq \sigma(T \cup \{w\}) - \sigma(T)$ whenever $S \subseteq T$. Informally, this is known as the "diminishing returns condition": the return derived from investing in node w diminishes as the size of the total investment (set) increases.

These properties of $\sigma(A)$ are sufficient to prove the desired approximation guarantee, for we can apply a well-known theorem of Nemhauser, Wolsey and Fischer. The first part of the theorem below is due to Nemhauser, Wolsey and Fischer [14, 15]; the generalization can be obtained by straightforward modifications to the proof.

Theorem 2. *Let f be a non-negative, monotone, submodular function on sets.*

1. *The greedy algorithm, which always picks the element v with largest marginal gain $f(S\cup\{v\})-f(S)$, is a $(1-1/\mathrm{e})$-approximation algorithm for maximizing f on k-element sets S.*
2. *A greedy algorithm which always picks an element v within $1-\varepsilon$ of the largest marginal gain results in a $1-1/\mathrm{e}-\varepsilon'$ approximation, for some ε' depending polynomially on ε.*

Given Theorem 2, in order to prove Theorem 1 (or its approximate version), it is sufficient to establish the following result:

Theorem 3. *For the decreasing cascade model, $\sigma(A)$ is a monotone and submodular function of A.*

Remark. The proof of the $(1 - 1/\mathrm{e})$ approximation guarantee in [6] was based on the same outline. In order to establish submodularity for the independent cascade and linear threshold models of [6], it was shown that for both models, it is possible to define distributions over directed graphs with the following property: for any set S of nodes, the probability that $\varphi(A) = S$ under the influence model is equal to the probability that the nodes of S are exactly the ones reachable from A in a graph chosen according to the corresponding distribution. Submodularity then follows readily from the fact that the number of reachable nodes in a fixed graph is a submodular function of the set of source nodes.

The decreasing cascade model is more general than the models considered in [6]. In Section 5, we give an instance which provably has no corresponding distribution on graphs. Therefore, the proof for submodularity becomes more intricate, and we have to consider the dynamics of the process in a more detailed way.

Most of the rest of this paper will be concerned with the proof of Theorem 3. We first introduce a generalized version of Granovetter's threshold model [10] in Section 3, as a useful reparametrization of the probability space. Using this threshold model, we then give the proof of Theorem 3 in Section 4.

3 The General Threshold Model

Recall that the notion of order-independence, as defined in Section 1.1, postulates that for a given set S of nodes trying to influence node v, the order in which these attempts are made does not affect the probability that v will be active once all the nodes in S have made their attempts. For the proof of Theorem 3, we require a stronger version of this statement: namely that even if the activation of nodes, or some activation attempts, are deferred for many time steps, the ultimate distribution over active sets remains the same.

It is not clear how to argue this fact directly from the definition of the cascade model, and we therefore introduce the general threshold model, a natural

generalization of Granovetter's linear threshold model [10]. The linear threshold model has been the foundation for a large body of work in sociology; see, e.g., [8, 16, 17, 18, 19, 20, 21]; its generalization was introduced in [6]. While the General threshold model is a natural model in its own right, in this work, we are most interested in it as a reparametrization of the cascade model. Indeed, Lemma 1 proves that the two models are equivalent.

In the *general threshold model* [6], each node v has a monotone *activation function* $f_v : 2^V \to [0, 1]$, and a threshold θ_v, chosen independently and uniformly at random from the interval $(0, 1]$. A node v becomes active at time $t + 1$ if $f_v(S) \geq \theta_v$, where S is the set of nodes active at time t. Again, the process starts with the activation of a select set A at time 1.

The threshold model focuses more on the "cumulative effect" of a node set S's influence on v, instead of the individual attempts of nodes $u \in S$. The perhaps somewhat surprising fact is that for any activation functions $f_v(\cdot)$, we can define corresponding success probabilities $p_v(\cdot, \cdot)$ such that the distribution over final active sets $\varphi(A)$ is identical under both models, for all sets A.

Specifically, given success probabilities $p_v(u, S)$, we define the activation functions

$$f_v(S) = 1 - \prod_{i=1}^{r} (1 - p_v(u_i, S_i)), \tag{1}$$

where $S = \{u_1, u_2, \ldots, u_r\}$, and $S_i = \{u_1, \ldots, u_{i-1}\}$. That f_v is well defined follows from the order-independence assumption on the $p_v(u, S)$. Conversely, given activation functions f_v, we define success probabilities

$$p_v(u, S) = \frac{f_v(S \cup \{u\}) - f_v(S)}{1 - f_v(S)}. \tag{2}$$

It is straightforward to verify that the activation functions defined via Equation (1) satisfy Equation (2), and the success probabilities defined via Equation (2) satisfy Equation (1).

Lemma 1. *Assume that the success probabilities $p_v(u, S)$ and activation functions $f_v(S)$ satisfy Equation (2). Then, for each node set T and each time t, the probability that exactly the nodes of set T are active at time t is the same under the order-independent cascade process with success probabilities $p_v(u, S)$ and the general threshold process with activation functions $f_v(S)$.*

Proof. We show, by induction, a slightly stronger statement: namely that for each time t and any pair (T, T'), the probability that exactly the nodes of T are active at time t, and exactly those of T' are active at time $t + 1$, is the same under both views. By summing over all sets T', this clearly implies the lemma.

At time $t = 0$, the inductive claim holds trivially, as the probability is 1 for the pair $(\emptyset, A)$ and 0 for all other pairs, for both processes. For the inductive step to time t, we first condition on the event that the nodes of T are active at time $t - 1$, and those of T' at time t.

Consider a node $v \notin T'$. Under the cascade process, v will become active at time $t+1$ with probability $1 - \prod_{i=1}^{r}(1 - p_v(u_i, T \cup T_i'))$, where we write $T' \setminus T = \{u_1, \ldots, u_r\}$ and $T_i' = \{u_1, \ldots, u_{i-1}\}$. Under the threshold process, node v becomes active at time $t+1$ iff $f_v(T) < \theta_v \leq f_v(T')$. Because node v is not active at time t, and by the Principle of Deferred Decisions, θ_v is uniformly distributed in $(f_v(T), 1]$ at time t, so the probability that v becomes active is $\frac{f_v(T') - f_v(T)}{1 - f_v(T)}$. Substituting Equation (1) for $f_v(T)$ and $f_v(T')$, a simple calculation shows that

$$\frac{f_v(T') - f_v(T)}{1 - f_v(T)} = 1 - \prod_{i=1}^{r}(1 - p_v(u_i, T \cup T_i')).$$

Thus, each individual node becomes active with the same probability under both processes. As both the thresholds θ_v and activation attempts are independent for distinct nodes, the probability for any set T'' to be the set of active nodes at time $t+1$ is the same under both processes. Finally, as the probability distribution over active sets T'' is the same conditioned on any pair (T, T') of previously active sets, the overall distribution over pairs (T', T'') is the same in both the cascade and threshold processes.

Lemma 1, which was stated without proof in [6], shows that the threshold model is a non-trivial reparametrization of the cascade model. In a natural way, it allows us to make all random choices at time 0, before the process starts. An alternate way of attempting to pre-flip all coins, for instance by providing a sequence of random numbers from $[0, 1]$ for use in deciding the success of activation attempts, would not preserve order-independence.

The nice thing about this view is that it makes a strong generalization of the notion of order-independence an almost trivial feature of the model. To formulate this generalization, we allow each node v a finite *waiting time* τ_v, meaning that when v's criterion for activation has been met at time t (i.e., an influence attempt was successful in the cascade model, or $f_v(S) \geq \theta_v$ in the threshold model), v only becomes active at time t+τ_v. Notice that when $\tau_v = 0$ for all nodes, this is the original threshold/cascade model.

Lemma 2. *Under the general threshold model, the distribution $\varphi(A)$ over active sets at the time of quiescence is the same regardless of the waiting times τ_v. This even holds conditioned upon any random event $\mathcal{E}$.*

Proof. We prove the stronger statement that for every choice of thresholds θ_v, and every vector τ of waiting times τ_v, the set S_τ of nodes active at the time of quiescence is the same as the set S_0 of nodes active at quiescence when all waiting times are 0. This will clearly imply the claim, by integrating over all thresholds that form the event $\mathcal{E}$. So from now on, fix the thresholds θ_v.

Let $A_{0,t}$ denote the set of nodes active at time t when all waiting times are 0, and $A_{\tau,t}$ the set of nodes active at time t with waiting times τ. A simple inductive proof using the monotonicity of the activation functions f_v shows that $A_{\tau,t} \subseteq A_{0,t}$ for all times t, which, by setting t to be the time of quiescence of the process with waiting times τ, implies that $S_\tau \subseteq S_0$.

Assume now that $S_\tau \neq S_0$, and let $T = S_0 \setminus S_\tau \neq \emptyset$. Among the nodes in T, let v be one that was activated earliest in the process without waiting times, i.e., $T \cap A_{0,t} = \emptyset$, and $v \in A_{0,t+1}$ for some time t. Because v was activated, we know that $\theta_v \leq f_v(A_{0,t})$, and by definition of v, no previously active nodes are in T, i.e., $A_{0,t} \subseteq S_\tau$. But then, the monotonicity of f_v implies that $\theta_v \leq f_v(S_\tau)$, so v should be active in the process with waiting times τ, a contradiction.

4 Proof of Theorem 3

The monotonicity is an immediate consequence of Lemma 3 below, applied with $V = V'$ and $p'_v(u, S) = p_v(u, S)$ for all S, v, u. So we focus on submodularity for the remainder of the proof. We have to show that, whenever $A \subseteq A'$, we have $\sigma(A \cup \{w\}) - \sigma(A) \geq \sigma(A' \cup \{w\}) - \sigma(A')$, for any node $w \notin A'$.

The basic idea of the proof is to characterize $\sigma(A \cup \{w\}) - \sigma(A)$ in terms of a *residual process* which targets only the node w, and has appropriately modified success probabilities (similarly for $\sigma(A' \cup \{w\}) - \sigma(A')$). To show that these residual processes indeed have the same distributions over final active sets $\varphi(\{w\})$ as the original processes, we use Lemma 2.

Given a node set B, we define the *residual process* on the set $V \setminus B$: the success probabilities are $p_v^{(B)}(u, S) := p_v(u, S \cup B)$, and the only node targeted is w, targeted at time 1. Let $\varphi_B(w)$ denote the set of nodes active at the time of quiescence of the residual process; notice that this is a random variable. We claim that, conditioned on the event that $[\varphi(A) = B]$, the variable $\varphi_B(w)$ has the same distribution as the variable $\varphi(A \cup \{w\}) \setminus \varphi(A)$.

In order to prove this fact, we focus on the threshold interpretation of the process, and assign node w a waiting time of $\tau_w = n + 1$. By Lemma 2, this view does not change the distribution of $\varphi(A \cup \{w\}) \setminus \varphi(A)$. Then, w is the only contagious node at time $n + 1$, and by the conditioning, the other active (but non-contagious) nodes are those from B. This implies that only nodes from $V \setminus B$ will make activation attempts after time $n + 1$. By using the same order of activation attempts, and the same coin flips for each pair $u, v \in V \setminus B$, a simple inductive proof on the time t shows that the set S of nodes is active in the residual process at time t if and only if the set $S \cup B$ is active in the original process at time $n + t$. In particular, this shows that the two random variables have the same distributions.

Having shown this equivalence, we want to compare the expected sizes of $\varphi_B(w)$ and $\varphi_{B'}(w)$, when $B \subseteq B'$. We write $\sigma_B(w) = \mathrm{E}[|\varphi_B(w)|]$, as well as $\sigma_{B'}(w) = \mathrm{E}[|\varphi_{B'}(w)|]$. First off, notice that the node set $V \setminus B$ of the former process is a superset of $V \setminus B'$. Furthermore, for all nodes u, v and node sets S, the decreasing cascade condition implies that

$$p_v^{(B)}(u, S) = p_v(u, S \cup B) \geq p_v(u, S \cup B') = p_v^{(B')}(u, S).$$

Lemma 3 below proves the intuitively obvious fact that the combination of a larger ground set of nodes and larger success probabilities results in a larger set of activated nodes, i.e.,

$$\sigma_w(B) \geq \sigma_w(B') \tag{3}$$

Finally, we can rewrite the expected number of active nodes as

$$\begin{aligned}
\sigma(A \cup \{w\}) - \sigma(A) &= \sum_B \sigma_w(B) \cdot \text{Prob}[\varphi(A) = B] \\
&= \sum_B \sum_{B' \supseteq B} \sigma_w(B) \cdot \text{Prob}[\varphi(A) = B, \varphi(A') = B'] \\
&\geq \sum_B \sum_{B' \supseteq B} \sigma_w(B') \cdot \text{Prob}[\varphi(A) = B, \varphi(A') = B'] \\
&= \sum_{B'} \sigma_w(B') \cdot \text{Prob}[\varphi(A') = B'] \\
&= \sigma(A' \cup \{w\}) - \sigma(A').
\end{aligned}$$

The inequality followed by applying Inequality (3) under the sum. In both of the steps surrounding the inequality, we used that $\text{Prob}[\varphi(A) = B, \varphi(A') = B'] = 0$ whenever $B \not\subseteq B'$, by the monotonicity of the cascade process. This completes the proof of submodularity. ∎

Lemma 3. *Let $V' \subseteq V$, and assume that $p'_v(u, S) \leq p_v(u, S)$ for all nodes $u, v \in V$ and all sets S. If $A' \subseteq A$ are the targeted sets for cascade processes on V' and V, then the expected size of the active set at the end of the process on V is no smaller than the corresponding expected size for the process on V'.*

Proof. This claim is most easily seen in the threshold view of the process. Equation (1) shows that the activation functions f'_v, f_v corresponding to the success probabilities $p'_v(u, S)$ and $p_v(u, S)$ satisfy $f'_v(S) \leq f_v(S)$, for all nodes v and sets S. Then, for any fixed thresholds θ_v, a simple inductive proof on time steps t shows that the set of active nodes in the former process (with functions f'_v) is always a subset of the set of active notes in the latter one (with functions f_v). Since the inequality thus holds for every point of the probability space, it holds in expectation.

5 Distributions over Graphs

As mentioned briefly before, the outline of the proof of the $(1 - 1/\text{e})$ approximation guarantee in [6] was the same as here. However, a simpler technique was used to show the submodularity of $\sigma(A)$.

This technique can be most easily understood in the case of the independent cascade model, where each activation attempt of a node u on a node v succeeds independently with probability $p_v(u)$. By the definition of the process, a node v is active in the end if it is reachable from one of the initially targeted nodes by a chain of successful activation attempts. If we consider a graph G that contains a directed arc (u, v) iff u's activation attempt on v succeeded, then it follows that a node v is active iff it is reachable in G from the targeted set

A. Due to the independence of activation attempts, and by the Principle of Deferred Decisions, the graph G can be generated by including each arc (u, v) independently with probability $p_v(u)$. As the set of nodes reachable from a given set A is a submodular function of A, and the expected size of the activated set is a non-negative linear combination (over all possible graphs G) of these functions, the function $\sigma(A)$ is shown to be submodular.

This technique can be applied whenever the influence model allows for a corresponding distribution on directed graphs G — the fact that we included each arc independently did not matter. In fact, [6] uses this technique to show submodularity in two other, less obvious, cases. In this section, we give an instance of the decreasing cascade model for which there is no distribution over graphs resulting in the same activation probabilities. This example shows that the techniques used to show submodularity of $\sigma(A)$ in [6] cannot be applied for the more general decreasing cascade model.

Our example has five nodes. Node v could potentially be influenced by four nodes $u_1, \ldots, u_4$. The first two nodes to try activating v have a probability of $\frac{1}{2}$ each to succeed, whereas all subsequent attempts fail. The influences are thus $p_v(u_i, S) = \frac{1}{2}$ whenever $|S| < 2$, and $p_v(u_i, S) = 0$ otherwise. Notice that this is indeed an instance of the decreasing cascade model, and order independent.

Assume, for contradiction, that there is a distribution on graphs such that node v is reachable from a set S with the same probability that S will activate v in the cascade model. For any set $S \subseteq \{1, 2, 3, 4\}$, let q_S denote the probability that in this distribution over graphs, exactly the edges from u_i to v for $i \in S$ are present. Because with probability $\frac{1}{4}$, v does not become active even if all u_i are, we know that $q_\emptyset = \frac{1}{4}$. If u_1, u_2, u_3 are active, then v is also active with probability $\frac{3}{4}$, so the edge (u_4, v) can never be present all by itself (if it were, then the set $\{u_1, u_2, u_3, u_4\}$ together would have higher probability of reaching v than the set $\{u_1, u_2, u_3\}$). Thus, we have that $q_{\{i\}} = 0$ for all i. The same argument shows that $q_{\{i,j\}} = 0$ for all i, j.

Thus, the only non-empty edge sets with non-zero probabilities can be those of size three or four. If node u_1 is the only active node, then v will become active with probability $\frac{1}{2}$, so the edge (u_1, v) is present with probability $\frac{1}{2}$. Hence, $q_{\{1,2,3\}} + q_{\{1,2,4\}} + q_{\{1,3,4\}} + q_{\{1,2,3,4\}} = \frac{1}{2}$, while $q_{\{1,2,3\}} + q_{\{1,2,4\}} + q_{\{1,3,4\}} + q_{\{2,3,4\}} + q_{\{1,2,3,4\}} = 1 - q_\emptyset = \frac{3}{4}$. Therefore, $q_{\{2,3,4\}} = \frac{1}{4}$, and a similar argument for nodes u_2, u_3, u_4 gives that $q_S = \frac{1}{4}$ for each set S of cardinality 3. But then, the total probability mass on edge sets is at least $\frac{5}{4}$, as there are four such sets S, and the empty set also has probability $\frac{1}{4}$. This is a contradiction, so there is no such distribution over graphs.

6 Conclusions

In this paper, we have presented and analyzed a simple greedy algorithm for maximizing the spread of influence in a general model of social influence termed the decreasing cascade model. The proof centered on showing that the expected number of influenced nodes is a monotone and submodular function of the tar-

geted set, which required new techniques beyond those used in previous work, including a non-trivial reparametrization of the probability space.

An interesting direction for future work is to investigate which are the most general influence models for which provable approximation guarantees can be achieved. A conjecture in [6], which is as of yet unresolved, states that whenever the activation functions f_v of the general threshold process of Section 3 are monotone and submodular at each node v, so is $\sigma(A)$.

Another direction for future work concerns the evaluation of the function $\sigma(A)$. At this point, we do not know if the function can be evaluated exactly in polynomial time, even for the simplest influence models.

References

1. Brown, J., Reinegen, P.: Social ties and word-of-mouth referral behavior. Journal of Consumer Research **14** (1987) 350–362
2. Domingos, P., Richardson, M.: Mining the network value of customers. In: Proc. 7th Intl. Conf. on Knowledge Discovery and Data Mining. (2001) 57–66
3. Goldenberg, J., Libai, B., Muller, E.: Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. Academy of Marketing Science Review (2001)
4. Goldenberg, J., Libai, B., Muller, E.: Talk of the network: A complex systems look at the underlying process of word-of-mouth. Marketing Letters **12** (2001) 211–223
5. Richardson, M., Domingos, P.: Mining knowledge-sharing sites for viral marketing. In: Proc. 8th Intl. Conf. on Knowledge Discovery and Data Mining. (2002) 61–70
6. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence in a social network. In: Proc. 9th Intl. Conf. on Knowledge Discovery and Data Mining. (2003) 137–146
7. Rogers, E.: Diffusion of innovations. 4th edn. Free Press (1995)
8. Valente, T.: Network Models of the Diffusion of Innovations. Hampton Press (1995)
9. Wasserman, S., Faust, K.: Social Network Analysis. Cambridge University Press (1994)
10. Granovetter, M.: Threshold models of collective behavior. American Journal of Sociology **83** (1978) 1420–1443
11. Berger, E.: Dynamic monopolies of constant size. Journal of Combinatorial Theory Series B **83** (2001) 191–200
12. Morris, S.: Contagion. Review of Economic Studies **67** (2000) 57–78
13. Peleg, D.: Local majority voting, small coalitions, and controlling monopolies in graphs: A review. In: 3rd Colloquium on Structural Information and Communication. (1996) 170–179
14. Cornuejols, G., Fisher, M., Nemhauser, G.: Location of bank accounts to optimize float. Management Science **23** (1977) 789–810
15. Nemhauser, G., Wolsey, L., Fisher, M.: An analysis of the approximations for maximizing submodular set functions. Mathematical Programming **14** (1978) 265–294
16. Macy, M.: Chains of cooperation: Threshold effects in collective action. American Sociological Review **56** (1991) 730–747

17. Macy, M., Willer, R.: From factors to actors: Computational sociology and agent-based modeling. Annual Review of Sociology **28** (2002) 143–166
18. Schelling, T.: Micromotives and Macrobehavior. Norton (1978)
19. Watts, D.: A simple model of fads and cascading failures. Technical Report 00-12-062, Santa Fe Institute Working Paper (2000)
20. Young, H.P.: Individual Strategy and Social Structure: An Evolutionary Theory of Institutions. Princeton University Press (1998)
21. Young, H.P.: The diffusion of innovations in social networks. Technical Report 02-14-018, Santa Fe Institute Working Paper (2002)