CT-IC: Continuously activated and Time-restricted Independent Cascade Model for Viral Marketing*

Wonyeol Lee, Jinha Kim, Hwanjo Yu#
Department of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
{*leewy, goldbar, hwanjoyu*}*@postech.ac.kr*

***Abstract*—Influence maximization problem with applications to viral marketing has gained much attention. Underlying influence diffusion models affect influence maximizing nodes because they focus on difference aspect of influence diffusion. Nevertheless, existing diffusion models overlook two important aspects of real-world marketing - *continuous trials* and *time restriction*. This paper proposes a new realistic influence diffusion model called *Continously activated and Time-restricted IC (CT-IC) model* which generalizes the IC model by embedding the above two aspects. We first prove that CT-IC model satisfies two crucial properties – *monotonicity* and *submodularity*. We then provide an efficient method for calculating *exact* influence spread when a social network is restricted to a directed tree and a simple path. Finally, we propose a scalable algorithm for influence maximization under CT-IC model called CT-IPA. Our experiments show that CT-IC model provides seeds of higher influence spread than IC model and CT-IPA is four orders of magnitude faster than the greedy algorithm while providing similar influence spread to the greedy algorithm.**

***Keywords*-influence maximization; viral marketing; social networks; influence diffusion model;**

I. Introduction

Due to the rapid growth of online social network sites such as Facebook and Twitter, we now experience that individuals' information is spread to others extremely fast. It enables us to use online social networks as a stage of viral marketing. However, when applying viral marketing, we face several important difficulties including *influence maximization problem*. Given a graph representing a social network, a parameter k denoting company's budget, and a stochastic process model of how influence is spread, the influence maximization problem aims at finding k *seeds* (initial nodes) which maximizes *influence spread*.

Kempe et al. [1] first propose the influence maximization problem and suggested two basic *influence diffusion models* – *Independent Cascade (IC) model* and *Linear Threshold (LT) model*. In IC model, an *active* node tries to activate its neighbors with a given probability and, in LT model, a node is activated only if some portion of its neighbors are already active. Recently, in [2], *IC model with negative opinions (IC-N)* is proposed, which considers the propagation of both negative and positive opinions.

Although several diffusion models have been suggested, they are too ideal to be applied to the real-world viral marketing applications. First, when a node becomes active, it can activate its neighbors *only once*. However, in real-world marketing situations, people influence his or her acquaintances *continuously*. Secondly, activation process is continued until *no more* activation happens at all. However, in the real world, we often have *time restriction* and thus cannot wait until the influence is spread "completely".

This paper proposes a more down-to-earth influence diffusion model for viral marketing applications called *Continuously activated and Time-restricted IC (CT-IC) model*. CT-IC model is a generalization of IC model, and it differs in two aspects: (a) every active node can activate its neighbors *repeatedly*, and (b) activations are processed until *a given time* T. Thus, CT-IC model requires two controllable parameters for the repeatable activations and the time constraint, and IC model is a special case of CT-IC model with a single activation and infinite time constraint.

After defining CT-IC model, we prove CT-IC model satisfies two crucial properties – monotonicity and submodularity – for influence spread, which proves that a simple greedy algorithm guarantees $(1-1/e)$-approximation under CT-IC model.

We then provide an efficient method for calculating *exact* influence spread when a graph is restricted to a directed tree. Because CT-IC model is a generalization of IC model, the equations computing the exact influence spread are more involved than those in IC model. We apply these equations to the special case of a directed tree, a simple path, to get a useful way to compute one node's influence on another node only through a path.

By using influence spread evaluation of a simple path, we propose an algorithm CT-IPA for CT-IC model which extends a scalable algorithm, *independent path algorithm* (IPA), for IC model [3]. Since influence spread of a *critical path* is computed by multiplying matrix weights of its edges, CT-IPA seamlessly extends IPA with additional treatments for merging multiple edges.

*This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7033344).

#corresponding author

1550-4786/12 $26.00 © 2012 IEEE
DOI 10.1109/ICDM.2012.40

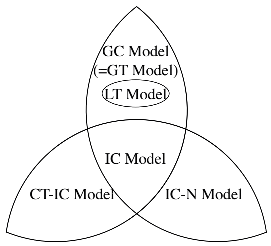


Figure 1. Relationship between influence diffusion models.

Experiments are conducted on two networks to find characteristic of CT-IC model and to compare CT-IPA with other algorithms. For the same dataset, CT-IC model and IC model produce seed sets of almost different nodes, and the nodes shared by two models have different ranks. In addition, CT-IPA shows over four orders of magnitude faster than greedy algorithm without sacrificing influence spread.

This paper is organized as follows. After describing related work in Section II, we propose CT-IC model and show its properties in Section III. Section IV presents efficient methods to compute exact influence spread. Section V proposes a scalable algorithm for influence maximization problem under CT-IC model. Section VI illustrates the experiment results, and Section VII concludes this paper. Full version of this paper is available at [4].

II. Related Work

Various influence diffusion models. Kempe et al. [1] suggest General Cascade (GC) model and General Threshold (GT) model which are generalized version of IC and LT models, and show that two models are equivalent. A new diffusion model called IC-N model, which considers the propagation of negative opinions, is recently proposed by Chen et al. [2]. However, all the above models restrict each node to have a single activation chance to its neighbors and do not consider time restriction of marketing.

The relationship between these diffusion models and CT-IC model is shown in Figure 1. All three diffusion models embed different aspects of influence propagation and have IC model as a common special case.

Efficient algorithms for influence maximization problem. Although the influence maximization problem is NP-Hard, Kempe et al. [1] show that the greedy algorithm guarantees $(1 - 1/e)$ approximation ratio under IC and LT models. It is shown by proving two properties of IC and LT models, *monotonicity* and *submodularity*. Mossel and Roch [5] and Chen et al. [2] prove that GC, GT models and IC-N model are also monotone and submodular, respectively.

However, the main drawback of the simple greedy algorithm is that it computes the influence spread by Monte Carlo simulation, which makes it too slow to be scalable. Thus, there have been many studies to reduce the running time of the original greedy algorithm. Several efficient algorithms are proposed based on approximating diffusion models such as Shortest Path Model [6], Prefix excluding Maximum Influence Arborescence (PMIA) [7], Local DAG [7], MIA with Negative opinions (MIA-N) [2], Community-based Greedy Algorithm [8], SIMPATH [9], IPA [3]. Also, Cost-Effective Lazy Forward (CELF) [10], CELF++ [11], and NewGreedy [12] are suggested by optimizing or changing the basic structure of greedy algorithm.

III. CT-IC Model

In this section, we describe the motivation of CT-IC model with several examples (Section III-A). Then, we formally define CT-IC model (Section III-B), and prove its important properties (Section III-C).

A. Motivation

Although IC model is widely used in data mining area, it is not realistic in two major aspects. Time limitation of marketing is ignored, and every node has a single chance of activation try. From now on, two motivational examples are provided to illustrate the importance of these two aspects.

First, let us consider a new release of iPhone 4s. After its release, most people will be interested in it and try to purchase it for a while. However, when another cutting edge iPhone 5 is released, most people will move their interest from iPhone 4s to iPhone 5. As a result, the amount of newly sold iPhone 4s would be very small compared to that before the release of iPhone 5. From this example, it is important to get maximum profit *within a time limit*.

Let us consider another example. Suppose you buy a new product and write a positive post about it in your Facebook wall. The post appears to your friends and persuades them to have a positive opinion, which may lead them to buy it. The important thing here is that when revisiting your wall later, your friends may be persuaded to buy the product although they were not persuaded before. In other words, your positive post will have *continuous* influence on your friends.

These two features – time constraint and continuous activation chances, which are not contained in IC model – are embedded in our new influence diffusion model.

B. Model definition

Let $G = (V, E)$ with a propagation probability $pp_0 : E \to (0, 1]$ be a directed graph representing a social network. $pp_0(u, v)$ denotes the probability that a node u activates a node v one time step after u is activated. Given a seed set $S \subseteq V$ and time restriction T, *Continuously activated and Time-restricted IC (CT-IC) model* works as follows.

First, every seed node $s \in S$ is activated at time 0, and the activation is propagated through its neighbors at time $t = 1, 2, 3, \cdots$. Let A_t be the set of active nodes at time t with $A_0 = S$. At time t, every active node $u \in A_t$ tries to activate its out-neighbors $v \in N_{out}(u)$ $(v \notin A_t)$ with probability $pp_{t-t_u}(u, v)$, where t_u is the activation time of u and $pp_t(u, v)$ is defined as

$$pp_t(u, v) = pp_0(u, v) \cdot f_{uv}(t).$$

Here, $f_{uv} : \mathbb{N}_0 \to \mathbb{R}_0^+$ is non-increasing and $f_{uv}(0) = 1$.

The non-increasing property of f_{uv} is based on the observation that persuading friends is getting harder after each trial to persuade them [13]. Accordingly, in this paper, we use $f_{uv}(x) = \exp(-\alpha_u x)$ with a non-negative constant α_u which represents how fast u's influence on its neighbors decreases.

After all activation trials are finished at time t, newly activated nodes S_t are included in the activated node set, so we have $A_{t+1} = A_t \cup S_t$ and a time step $t+1$ starts. This activation process is repeated until we arrive at time step T.

The big difference between CT-IC and IC model is that (a) every active node has *multiple* chances to activate its neighbors (until its neighbor becomes active), not only right after it is activated, and (b) *all* activation processes *stop* at global time T, not at time ∞.

The influence spread of a given seed set S at time t, $\sigma(S,t)$, is the expected number of active nodes when time step t starts. Then, given the number of seed nodes k and time constraint T, the influence maximization problem under CT-IC model is to find a set $S^* \in \arg\max_{S \subseteq V, |S|=k} \sigma(S,T)$.

CT-IC model is a generalized version of IC model because IC model is obtained by taking $T = |V|, \alpha_v \to \infty, \forall v \in V$.

C. Properties of CT-IC model

Monotonicity and submodularity. In order to ensure that greedy algorithm produces $(1-1/e)$-approximation solution for influence maximization problem under CT-IC model, monotonicity and submodularity of CT-IC model should be proven. Here, for a given function $f : 2^V \to \mathbb{R}$, f is called *monotone* if $f(S) \leq f(S')$, $\forall S \subseteq S'$, and *submodular* if $f(S \cup \{v\}) - f(S) \geq f(S' \cup \{v\}) - f(S')$, $\forall S \subseteq S', v \in V$.

To prove monotonicity and submodularity, we conceive an easy-to-analyze process which is equivalent to CT-IC model. Consider an edge $(u,v) \in E$. After u is newly activated at t_u, u tries to activate v continuously until v becomes active. Then, the probability that v is activated exactly at $t_u + t$ by u is $pp_{t-1}(u,v) \prod_{i=0}^{t-2}(1 - pp_i(u,v))$. Since the probability that u activates v at $t_u + t$ for each $(u,v) \in E$ is given as above, we can decide t before activation process starts, and it is an equivalent activation process to CT-IC model.

Suppose that we decide t for each $(u,v) \in E$ before activation process starts, and represent it as a function $h : E \to \mathbb{N}$. Let $G' = (V,E,h)$ be a graph with weight $h(u,v)$ for $(u,v) \in E$. Then, it is easily shown that $v \in V$ is active at time t if and only if there exists $u \in S$ and a path from u to v in G' whose length is equal to or less than t, where S is a seed set. Based on this observation, after choosing h, we can compute influence spread deterministically (Theorem 1).

Theorem 1: $\sigma(\cdot,t)$ is monotone and submodular, $\forall t \geq 0$.

Proof: See [4]. ■

As $\sigma(\cdot,t)$ under CT-IC model is monotone and submodular by Theorem 1 and it is trivially non-negative, Greedy algorithm (Algorithm 1) guarantees a $(1-1/e)$-approximation solution by Theorem 2.1 in [1].

Algorithm 1 Greedy(G,k,T)

$S = \phi$
for $i = 1$ **to** k **do**
  $u = \arg\max_{v \in V \setminus S} \sigma(S \cup \{v\}, T) - \sigma(S,T)$
  $S = S \cup \{u\}$
end for
return S

Difference between IC and CT-IC models. To investigate how different CT-IC model is from IC model, we now introduce a measure called *difference ratio between IC and CT-IC model* as follows. Assume that $G = (V,E)$, k, and T are given. Define the set of optimal solutions for CT-IC model and that of IC model as $\mathfrak{S}_I^*(G,k) = \arg\max\{\sigma_I(S) | S \subseteq V, |S| = k\}$, $\mathfrak{S}_T^*(G,k) = \arg\max\{\sigma(S,T) | S \subseteq V, |S| = k\}$, respectively, where $\sigma_I(S)$ is the influence spread of seed set S in IC model. Then, we define the difference ratio as

$$dr(G,k,T) = \frac{\sigma(S_T^*, T)}{\max\{\sigma(S_I^*, T) | S_I^* \in \mathfrak{S}_I^*\}} \geq 1,$$

where $S_T^* \in \mathfrak{S}_T^*$. dr tells that whether we can get good solution for influence maximization under CT-IC model even if we just treat CT-IC model as IC model. This ratio can be used as a measure to quantify the difference between IC and CT-IC models.

The following Lemma says that for small k, T, there exist infinitely many graphs for which dr is sufficiently large.

Lemma 1: For any positive k, N, T such that $k < N/4$, $T < (N/4k) - 1 = O(N/k)$, there exists a graph $G = (V,E)$ such that $|V| = N$ and $dr(G,k,T) = \Omega(N/kT)$.

Proof: See [4]. ■

IV. Exact computation of influence spread

Because computing influence spread under IC model is #P-Hard [14] and IC model is a special case of CT-IC model, computing influence spread under CT-IC model is also #P-Hard. However, its computation is still tractable if we restrict the whole graph to an arborescence (Section IV-A) or to a simple path (Section IV-B).

A. Case of an arborescence

Consider an arborescence $G_A = (V,E)$ with a seed set $S \subseteq V$ and time restriction T. For any $v \in V$ and $0 \leq t \leq T$, let $ap_S(v,t)$ be a probability that v is activated *exactly at* time t, and $ap_{S,T}(v)$ be a probability that v is activated before activation process ends (i.e. $ap_{S,T}(v) = \sum_{i=0}^{T} ap_S(v,i)$). Then, it is obvious that

$$ap_S(v,t) = \begin{cases} 1 & \text{if } v \in S \text{ and } t = 0 \\ 0 & \text{if } v \notin S \text{ and } t = 0 \\ 0 & \text{if } v \in S \text{ and } t > 0 \end{cases}.$$

However, when $v \notin S$ and $0 < t \leq T$, computing $ap_S(v,t)$ is not trivial. The following Lemma 2 tells that in this case, $ap_S(v,t)$ has a complex formula.

Lemma 2: For any $v \in V \backslash S$ and $0 < t \leq T$,

$$ap_S(v,t) = \prod_{u \in N_{in}(v)} \left[1 - \sum_{i=0}^{t-2} ap_S(u,i) g_{uv}(t-2-i)\right] - \prod_{u \in N_{in}(v)} \left[1 - \sum_{i=0}^{t-1} ap_S(u,i) g_{uv}(t-1-i)\right]$$

holds, where $g_{uv}(t) = 1 - \prod_{i=0}^{t} [1 - pp_i(u,v)]$.

Proof: See [4]. ■

We know that $\sigma(S,T) = ap_{S,T}(v)$ holds. Therefore, when a given graph is an arborescence, we can compute the exact value of $\sigma(S,T)$ in $O(|V|T^2)$ time by using dynamic programming.

B. Case of a simple path

Let us consider a general directed graph $G = (V,E)$ with seed set $S \subseteq V$ and time restriction T. Let $v \in V \backslash S$ and $u \in N_{in}(v)$. To find out an equation which evaluates influence spread along a simple path, assume that v is activated *only* by u. Then, by using Lemma 2, $ap(v,t)$ is calculated as

$$ap(v,t) = \sum_{i=0}^{t-1} c_{uv}^{(t-i)} ap(u,i) = \begin{bmatrix} ap(u,0) \\ ap(u,1) \\ \vdots \\ ap(u,t-1) \end{bmatrix}^{\mathrm{Tr}} \begin{bmatrix} c_{uv}^{(t)} \\ c_{uv}^{(t-1)} \\ \vdots \\ c_{uv}^{(1)} \end{bmatrix},$$

where $c_{uv}^{(t-i)} = pp_{t-i-1}(u,v) \prod_{j=0}^{t-i-2} (1 - pp_j(u,v))$. Obvious subscript S in $ap_S(v,t)$ is omitted. After putting $ap(v,i)$'s for $i = 0, \cdots, T$ into a matrix, we have

$$\begin{bmatrix} ap(v,0) \\ ap(v,1) \\ ap(v,2) \\ \vdots \\ ap(v,T) \end{bmatrix}^{\mathrm{Tr}} = \begin{bmatrix} ap(u,0) \\ ap(u,1) \\ ap(u,2) \\ \vdots \\ ap(u,T) \end{bmatrix}^{\mathrm{Tr}} \begin{bmatrix} 0 & c_{uv}^{(1)} & \cdots & c_{uv}^{(T)} \\ 0 & 0 & \cdots & c_{uv}^{(T-1)} \\ 0 & 0 & \cdots & c_{uv}^{(T-2)} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix},$$

or $\mathbf{AP}(v) = \mathbf{AP}(u)\mathbf{C}_{uv}$ equivalently, where $\mathbf{AP}(v)$, $\mathbf{AP}(u)$, and $\mathbf{C}_{uv}$ represent corresponding matrices.

Now, for any $u \in S$ and $v \in V \backslash S$, consider a simple path $p = (u = u_0, u_1, \cdots, u_{l-1}, u_l = v)$ where $u_i \in V \backslash S$ for all $i = 1, \cdots, l$, let us define $inf_p(u,v)$ be the probability that u activates v until time T only through a path p. By using the above result, we have $\mathbf{AP}(v) = \mathbf{AP}(u)\mathbf{C}_{u_0 u_1} \cdots \mathbf{C}_{u_{l-1} u_l}$. However, we know that $\mathbf{AP}(u) = [1\ 0\ \cdots\ 0]$ and $ap_{S,T}(v) = \sum_{i=0}^{T} ap_S(v,i) = \mathbf{AP}(v)[1 \cdots 1]^{\mathrm{Tr}}$. Therefore, we finally obtain the following formula.

$$inf_p(u,v) = [1\ 0\ \cdots\ 0] \left(\prod_{i=0}^{l-1} \mathbf{C}_{u_i u_{i+1}} \right) [1\ 1\ \cdots\ 1]^{\mathrm{Tr}}, \quad (1)$$

In IC and IC-N models, the property that all sub-paths of any maximum probability path are also maximum probability paths holds, so we could make a reasonable local tree structure, such as MIA [7] and MIA-N [2], for efficient algorithms. However, CT-IC model does not have such

Table I
BASIC INFORMATION OF FOUR REAL DATASET.

Dataset	Directedness	$\lvert V \rvert$	$\lvert E \rvert$	θ for CT-IPA
HEP	Undir	$15K$	$59K$	$1/32$
AMAZON	DIR	$262K$	$1235K$	$1/16$

property. (See Lemma 4 in [4].) Therefore, obtaining similar local arborescences of MIA or MIA-N for CT-IC model is computationally intractable because shortest path algorithm such as Dijkstra's algorithm cannot be used for finding maximum probability path.

V. Influence Spread Processing Algorithm

In this section, we propose *Continuously activated and Time-restricted influence path algorithm* (CT-IPA) for CT-IC model by extending a highly scalable algorithm for IC model – independent path algorithm (IPA) [3]. The two assumptions IPA is based on are that influence is propagated only through *critical paths* (paths whose influence spread is greater than a threshold θ) and activation process through each critical path is independent of each other.

The extension from IPA to CT-IPA is seamlessly done by changing the influence spread definition of an influence path. In CT-IC model, influence spread of an influence path is $inf_p(\cdot,\cdot)$ of Equation 1 which involves matrix multiplication. Therefore, embedding $inf_p(\cdot,\cdot)$ into IPA, we get CT-IPA algorithm for CT-IC model. For detailed description of IPA and CT-IPA, please refer to [3] and [4].

VI. Experiments

In this section, we conduct experiments to figure out characteristic of CT-IC model and to compare the performance of CT-IPA with other algorithms. See [4] for more experiment results.

A. Experiment Setup

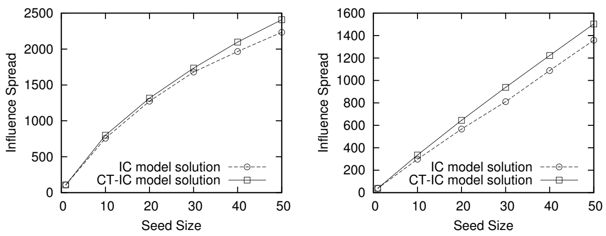
Datasets. Two real datasets are used in experiments. HEP is a co-authorship graph obtained from "High Energy Physics - Theory" section of arXiv site (http://arxiv.org) where nodes and edges represent authors and coauthor relationships. AMAZON is a co-purchasing graph of amazon.com, in which a node u represents a product and an edge (u,v) represents that v is usually bought with u. We get HEP data from Wei Chen's site[1], and AMAZON from Jure Leskovec's site[2]. The basic statistics of each graph is presented in Table I.

Propagation probabilities. WC (weighted cascade) model [1] is used for generating edges' probabilities. In WC model, propagation probabilities are assigned as $pp_0(u,v) = 1/\deg_{in}(v)$ for all edges $(u,v) \in E$, where $\deg_{in}(v)$ denotes the in-degree of node u.

Algorithms.

- Random : A baseline algorithm which selects k nodes uniformly at random from the overall $|V|$ nodes.

[1] http://research.microsoft.com/en-us/people/weic/graphdata.zip
[2] http://snap.stanford.edu/data



(a) HEP (b) AMAZON

Figure 2. Comparison between IC and CT-IC models.

Table II
TOP-20 SEED NODES OF IC MODEL AND CT-IC MODEL SOLUTION.

(a) On HEP

IC model solution	639	131	100	124	287
	562	606	15	382	196
	267	608	1162	66	274
	307	4824	128	989	412
CT-IC model solution	639	124	131	100	66
	274	287	606	412	**236**
	1162	608	**221**	**1292**	**80**
	192	267	**76**	**239**	4824

(b) On AMAZON

IC model solution	17747	222839	25699	18076	168039
	18337	232448	7266	11129	45391
	176067	9657	64815	183084	27562
	59541	14461	238375	114241	1385
CT-IC model solution	17747	176067	**56415**	**51234**	**200657**
	238375	18076	**236670**	**259011**	222839
	6290	**205434**	**143531**	**199539**	59541
	25699	**178335**	**82533**	114241	**95315**

- MaxDegree : A simple heuristic which selects k nodes in non-increasing order of node's out-degree.
- Greedy : Algorithm 1 with lazy-forward optimization [10]. We use $10,000$ times of Monte-Carlo simulations to compute $\sigma(S, T)$.
- CT-IPA : Our proposed algorithm integrated with lazy-forward greedy optimization. The last column of Table I shows tuned θ values used on each dataset.[3]

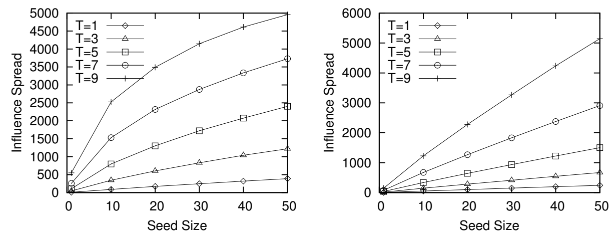
We do not include any algorithms for IC model because they are not extendable to CT-IC model as described in Section IV-B.

In the experiment, we set $\alpha_v = 0.1$ for all v. Different α values produced similar results. When we calculate the influence spread of each seed set produced by each algorithm, we do $10,000$ Monte-Carlo simulations and get the average of the values. We conduct the following experiments in a Linux machine with two Intel Xeon CPUs and 24GB memory.

B. Characteristic of CT-IC model

Comparison between IC and CT-IC models. In order to check whether CT-IC model is really different from IC model, we compare the greedy algorithm for "IC model" and the greedy algorithm for "CT-IC model" . After obtaining solutions (seed sets) from two methods, we calculate the influence spread of them under **"CT-IC model"** and compare it. Since it is not feasible to get a solution by greedy algorithms for AMAZON, we use IPA [3] and CT-IPA instead of greedy algorithm for IC and CT-IC models, respectively. We vary seed size k from 1 to 50, and set $T = 5$.

[3] We find that there is trade-off between processing time and influence spread as θ changes. Thus, by varying $\theta = 1/8, 1/16/, \cdots, 1/512$, we select the first θ at which an increment in influence spread becomes much smaller than that in processing time.



(a) HEP (b) AMAZON

Figure 3. The change of influence spread with respect to T.

Figure 2 shows the influence spread of two methods' solutions on four datasets. On both datasets, the influence spread of CT-IC model solution is always larger than that of IC model solution, and moreover the gap between them becomes larger as k increases. These results show that (1) CT-IC model is a different model from IC model and (2) time constraint and continuous activation trials of CT-IC model are meaningful consideration for a realistic influence diffusion model.

In addition to influence spread comparison, we compare top-20 node solutions for IC and CT-IC models. The results on HEP and AMAZON are listed in Table II. In the node id list, the top-left node is top-1st node of solution and the bottom-right node is the top-20th node, and node ids in bold type are ones which are included in CT-IC model solution but not in IC model solution.

Among top-20 nodes, only 13 and 6 nodes are in common for both solutions on HEP and AMAZON, respectively. Moreover, the ranking of top-20 nodes in CT-IC model solution is largely different from that in IC model solution. Thus, CT-IC model is a more different model from IC model than it appears in Figure 2.

Change of influence spread when varying T. To find out how influence spread changes as T increases, influence spread is measured when $T = 1, 3, \cdots, 9$. k is also varied from 1 to 50. We select seed nodes by Greedy for HEP and by CT-IPA for AMAZON.

Figure 3 illustrates the results of influence spread. On every dataset, influence spread increases as T increases, which is obvious. However, on HEP, as T increases, the increment of influence spread *increases* at first, and then starts to *decrease* at some point. The fact that the increment of influence spread *increases* is not intuitive but can be explained as follows.

Let $\Delta[\sigma](S, T) = \sigma(S, T+1) - \sigma(S, T)$ and $\Delta^2[\sigma](S, T)$

$= \Delta[\sigma](S, T+1) - \Delta[\sigma](S, T)$. The above observation is almost equivalent to that "$\Delta^2[\sigma](S, T)$ is at first positive but becomes negative at some point as T increases." There are two opposite effects on the sign of $\Delta^2[\sigma]$ – the effects of *already* active nodes and *newly* activated nodes. The nodes that are already active at T activate less nodes as time goes by because pp_t keep decreasing, and try to make $\Delta^2[\sigma]$ negative. On the other hand, because the nodes that are newly activated at $T+1$ are not active before $T+1$, their activation tries only increase $\Delta[\sigma](S, T+1)$ and make it positive. By this argument, we can now explain the above observation – $\Delta^2[\sigma] < 0$ (resp. > 0) because the first effect (resp. the second one) is stronger than the other.

C. Comparison between algorithms

Influence spread. We measure the influence spread of algorithms' solutions by varying k from 1 to 50. We set $T = 5$. Greedy is only applied to HEP because of its excessive processing time on AMAZON. The results are shown in Figure 4.

On HEP, the influence spread of CT-IPA is quite close to that of Greedy but there is a significant gap between CT-IPA and MaxDegree. On AMAZON, CT-IPA is still overwhelmingly the best, and the influence spread of CT-IPA is almost linear to k, like in IC model [7].

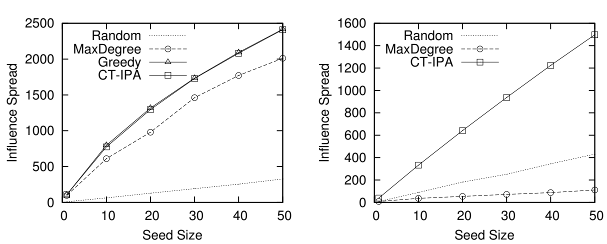
In a nutshell, CT-IPA yields influence spread as high as Greedy, and always shows better influence spread than other algorithms. Additionally, MaxDegree is very unstable. Though it performs well in few cases, it does not in other cases and is sometimes worse than Random.

Processing time. We measure the processing time of algorithms on two datasets up-to 10 hours when $k = 50$ and $T = 5$. The result is shown in Figure 5.

The processing time of Greedy is 5.0 hours on HEP, and more than 10 hours on AMAZON. Thus, as in IC model, Greedy is absolutely not scalable. On the other hand, CT-IPA runs only in 7.0, and 14.3 seconds on each graph, which is 4 orders of magnitude faster than Greedy. Since MaxDegree and Random are very simple, they always take less than one second. However, influence spread of their solutions is unstable and much worse than CT-IPA.

VII. Conclusion

In this paper, we propose a realistic influence diffusion model – the Continuously activated and Time-restricted independent cascade (CT-IC) model. Existing influence diffusion models and their efficient processing algorithms lack of two important aspects of influence propagation in real world – time constraint and continuous activation trials. CT-IC model embeds these two aspects into its activation process. By proving monotonicity and submodularity, the greedy algorithm which has $1-1/e$ approximation ratio can be applied to CT-IC model. Moreover, exact influence spread evaluation in CT-IC for a specific graph (e.g. arborescences and simple paths) are derived. By plugging the exact influence spread evaluation of simple paths to IPA algorithm for IC model, we have a highly scalable processing algorithm CT-IPA for CT-IC model. Extensive experiments on real datasets show that CT-IC model produces different results from IC model, and CT-IPA produces seed sets several orders of magnitude faster than the greedy algorithm without sacrificing influence spread.



(a) HEP (b) AMAZON

Figure 4. Influence spread of various algorithms.

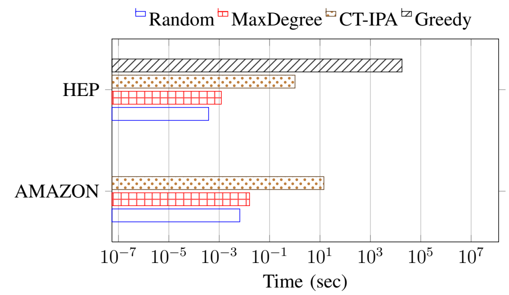


Figure 5. Processing time of various algorithms.

References

[1] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *KDD 2003*.

[2] W. Chen, A. Collins, R. Cummings, T. Ke, Z. Liu, D. Rincón, X. Sun, Y. Wang, W. Wei, and Y. Yuan, "Influence maximization in social networks when negative opinions may emerge and propagate," in *SDM 2011*.

[3] J. Kim, S.-K. Kim, and H. Yu, "Scalable and parallelizable processing of influence maximization for large-scale social network," Pohang University of Science and Technology (POSTECH), Tech. Rep., 2012. [Online]. Available: http://dm.postech.ac.kr/techreport/TechReport-POSTECH-CSE-2012-02-IPA.pdf

[4] W. Lee, J. Kim, and H. Yu, "Ct-ic: Continuously activated and time-restricted independent cascade model for viral marketing," Pohang University of Science and Technology (POSTECH), Tech. Rep., 2012. [Online]. Available: http://dm.postech.ac.kr/techreport/TechReport-POSTECH-CSE-2012-02-CT-IC.pdf

[5] E. Mossel and S. Roch, "On the submodularity of influence in social networks," in *STOC 2007*.

[6] M. Kimura and K. Saito, "Tractable models for information diffusion in social networks," in *ECML PKDD 2006*.

[7] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *KDD 2010*.

[8] Y. Wang, G. Cong, G. Song, and K. Xie, "Community-based greedy algorithm for mining top-k influential nodes in mobile social networks," in *KDD 2010*.

[9] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Simpath: An efficient algorithm for influence maximization under the linear threshold model," in *ICDM 2011*.

[10] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *KDD 2007*.

[11] A. Goyal, W. Lu, and L. V. S. Lakshmanan, "Celf++: optimizing the greedy algorithm for influence maximization in social networks," in *WWW 2011*.

[12] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *KDD 2009*.

[13] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *WSDM 2010*.

[14] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *ICDM 2010*.