A Unified Influence Maximization Processing Framework for Independent Cascade Model and Its Extensions *

Jinha Kim, Seung-Keol Kim, Hwanjo Yu†
Department of Computer Science and Engineering
Pohang University of Science and Technology (POSTECH)
Pohang, South Korea
{goldbar,azzibobj,hwanjoyu}@postech.ac.kr

ABSTRACT

The *"word-of-mouth"* effect on social networks – opinion propagation – nowadays plays an important role in marketing. The influence maximization problem aims at maximizing the word-of-mouth effect and formulates such goal as a combinatorial optimization problem. As an important process of the influence maximization problem, evaluating influence is #P-hard which cannot be affordable in a polynomial time. Accordingly, devising a fast influence evaluation/approximation algorithm is crucial. We propose a scalable influence approximation algorithm called IPA. The main idea of IPA is considering an influence path as an independent influence evaluation unit. Using that idea, IPA approximates influence an order of magnitude faster and more accurately, and uses much less memory than PMIA, the current state of the art approximation algorithm. In addition, IPA is applicable to independent cascading (IC) model and its extensions (IC-N, t-IC, t-IC-N models), whereas PMIA is only adaptable to IC and IC-N models. We implement and demonstrate a desktop application which finds the most influential nodes and visualizes influence paths. The application is accessible at `http://dm.postech.ac.kr/ipa_demo`.

Categories and Subject Descriptors

F.2.2 [**Analysis of Algorithms and Problem Complexity**]: Non-numerical Algorithms and Problems

General Terms

Algorithms, Experimentation, Performance

Keywords

Influence maximization, social networks

*This work was partially supported by the Brain Korea 21 Project in 2012 and Mid-career Researcher Program through NRF grand funded by the MEST (No. KRF-2011-0016029). This work was also supported by IT Consilience Program of MKE and NIPA (C1515-1221-003).

†corresponding author

1. INTRODUCTION

As online social networks such as Twitter and Facebook are rapidly growing, our opinions on products or political topics propagate fast to other users through the networks, which is called *"word-of-mouth"* effect. Accordingly, corporations start to treat social networks as a stage of viral marketing by exploiting the word-of-mouth effect. To maximize effect of viral marketing, researchers in data-mining community introduce the influence maximization problem [6] which addresses the most influential users in a social network.

Even though the influence maximization problem has effectiveness and scalability challenges, sophisticated methods detour these challenges. To overcome NP-hardness, greedy algorithm [6] approximates the optimal solution with lower bound ratio of $(1 - 1/e)$. To overcome #P-hardness of evaluating influence which dominates the processing time, existing works exploit smaller communities [12], shortest path [9], and local influence structure [2, 3].

Along with effective and efficient processing algorithms, new influence diffusion models are suggested which are more realistic than actively studied models like independent cascade (IC) and linear threshold (LT) models. IC model with negative opinion (IC-N) [4] embeds negative opinion on a marketed product. Time-considering IC (t-IC) model [10] adds time limit constraint of viral marketing and continuous trial of influencing neighbors. Also, t-IC-N is the mixture of IC-N and t-IC model.

In this paper, we propose an influence spread approximation algorithm, IPA, which is (1) highly scalable for both processing time and memory space perspectives and (2) extendable to any kind of IC model. By considering an influence path as an independent influence evaluation unit, IPA simplifies influence spread evaluation. Accordingly, the processing time of IPA is an order of magnitude faster than that of PMIA [2] without sacrificing influence spread much. IPA also reduces memory usage by throwing away insignificant influence paths safely. In addition, the idea of independent influence evaluation unit makes IPA applicable to any extension of IC model. Although PMIA aims at IC model and applicable to IC-N model , it is not adaptable to time-considering models such as t-IC and t-IC-N models. However, because IPA catches the essence of IC model – independent one-to-one persuasion, it is adaptable to any kind of IC model by defining the influence propagation probability under each model. Moreover, the simple structure of IPA enables parallel processing, and the implementation of parallel IPA requires only a few lines of OpenMP [5] meta-programming expressions.

We also present a desktop application which implements IPA for IC, IC-N, t-IC, and t-IC-N models. The application demonstrates that for million-nodes graph IPA provides the most influential nodes within ten minutes without losing influence spread much.

2. IC MODEL AND ITS EXTENSIONS

Influence diffusion models reflect and simplify influence dynamics of social networks. Because real influence dynamics of social networks is too complex, each influence diffusion model focuses on a specific aspect. As one of the representative diffusion model, independent cascade (IC) model [6] assumes that an influence propagation happens by one-to-one persuasion. Due to its simplicity, IC model is the most widely used in various literature [1, 2, 6, 11, 12]. In this paper, we will provide solution of influence maximization under IC model and its extended models.

When a directed graph $G(V, E)$ is an abstracted social network where $v \in V$ is a node representing a user and $(u, v) \in E$ is an edge representing relationship between users, the dynamics of IC model works in an inductive way. Each node has one of two state - *active*(seed or influenced) or *inactive*(not yet influenced) and each edge $(u, v) \in E$ has a constant propagation probability $w_{(u,v)}$. Let A_t denote a set of nodes which become active at stage t and $N_{in}(v)$ denote the in-neighbor set of a node v. At initial stage 0, $S \subseteq V$ is chosen as a seed set and $A_0 = S$. At stage $t+1$, each inactive node $v \notin \bigcup_{i=0}^{t} A_i$ has a chance to be influenced by v's in-neighbor $u \in A_t \cap N_{in}(v)$ with the probability of $w_{(u,v)}$. The influence diffusion ends when $A_t = \phi$.

IC model with negative opinion (IC-N) [4] is an extension of IC model which considers negative opinion on a marketed product. Unlike IC model, each node has one of three states - *positive*, *negative*, or *inactive*. The active state of IC model is divided into positive or negative state. The dynamics of IC-N model is slightly different from that of IC model due to negative opinion handling. Let $P_t(N_t)$ denote a set of positive(negative) nodes activated at stage t, and q denote quality factor – the probability of keeping positive opinion when a node is activated. At stage 0, each seed node $s \in S$ belongs to P_0 with the probability q, otherwise belongs to N_0. At stage $t+1$, for an inactive node v, v's active neighbor $u \in N_{(in)}(v) \cap (P_t \cup N_t)$ tries to influence v. When u is positive, v becomes positive(negative) with the probability $q \cdot w_{(u,v)}((1-q) \cdot w_{(u,v)})$. When u is negative, v always becomes negative with the probability $w_{(u,v)}$. To determine which $u \in N_{(in)}(v) \cap (P_t \cup N_t)$ influences v, we first generate a random permutation and influence trials follow the order of the permutation.

Time-considering IC (t-IC) model [10] is another extension of IC model which embeds two aspects of viral marketing of the real world; (1) time limit of viral marketing and (2) continuous influence trial. Let T denote the marketing ending time, and t_v denote the stage when a node v becomes active. For the time limit, influence diffusion ends at T, even though $A_T \neq \phi$. For the continuous trial, a node u which becomes active at t_u tries to activate its inactive out-neighbor v until T with the probability of $w_{(u,v)} \cdot \delta(t, t_u)$. $\delta(t, t_u)$ is a non-increasing decaying function which represents the diminishing influence probability of u to v. Typically, $\delta(t, t_u) = \exp^{-\alpha(t-t_u)}$ where α is a parameter that controls the decaying speed.

As the most complex extension of IC model, time-considering IC model with negative opinion (t-IC-N) can also be considered. In t-IC-N model, negative opinion possibly emerge during influence diffusion, and the viral marketing has its ending time and each influenced node continuously tries to its inactive out-neighbors with diminishing probability until the marketing ends. The dynamics of t-IC-N model is the mixture of IC-N model and t-IC model. Also, t-IC-N model has all the previous models as its special cases. IC-N model is an instance of t-IC-N model with $T = \infty$ and $\alpha = \infty$. t-IC model is an instance of t-IC-N model with $q = 1$. IC model is instance of t-IC-N model with $T = \infty, \alpha = \infty$ and $q = 1$.

3. INFLUENCE MAXIMIZATION PROBLEM

Given a graph and an influence diffusion model, the influence maximization problem must be formalized to solve it in an algorithmic way. Kempe et al. [6] first formalize the influence maximization problem as a combinatorial optimization problem and their formulation has been actively researched [1, 2, 6, 7, 9, 11, 12]. We also follow the formulation of [6].

With a directed graph G, the influence maximization problem is formulated as follows:

Definition 1. The influence maximization problem addresses the top-k nodes which satisfies

$$S = \arg \max_{T \subseteq V, |T|=k} \sigma(T) \tag{1}$$

where $\sigma(T)$ returns the expected number of activated node from a node set T and is called *influence spread*.

The solution of the influence maximization differs by the underlying influence diffusion model. Influence diffusion model determines which nodes are activated, and consequently generates different value of influence spread even for the same seed set. From now on, we call influence spread of each influence diffusion model as follows. $\sigma_I(S)$ is influence spread of S under IC model. $\sigma_N(S, q)$ is positive influence spread with quality factor q under IC-N model. $\sigma_t(S, T, \alpha)$ is influence spread with time limit T and decaying factor α under t-IC model. $\sigma_{tN}(S, q, T, \alpha)$ is positive influence spread under t-IC-N model.

Greedy Algorithm. One challenge of the influence maximization problem of Definition 1 is its NP-hardness [6]. To find the optimal solution, the search space of Definition 1 is ${}_{|V|}C_k = \frac{|V|!}{k!(|V|-k)!} \approx |V|^k$ (usually $k \ll |V|$), which is exponential in terms of k. The NP-hardness of IC model is proved in [6], and IC-N, t-IC, t-IC-N models are also NP-hard because they have IC model as their special case.

To detour NP-hardness of the influence maximization problem, Kempe et al. [6] approximate the optimal solution using greedy hill-climbing approach called Greedy. Its approximation ratio to the optimal solution is $1 - 1/e \approx 0.631$. CELF greedy algorithm

Algorithm 1 Greedy(k, f)

1: $S \leftarrow \phi$
2: **for** $i = 1$ to k **do**
3: $\quad v \leftarrow \arg\max_{u \in V \setminus S} f(S \cup \{u\}) - f(S)$
4: $\quad S \leftarrow S \cup \{v\}$
5: **end for**
6: **return** S

[11] efficiently finds v of line 3 in Algorithm 3 by using lazy-forward evaluation.

To apply Greedy to the influence maximization problem, the optimization target function $\sigma(S)$ must satisfy three properties; non-negativity, monotonicity, and submodularity. There three properties are proved under all the influence diffusion model introduce in Section 2 – $\sigma_I(S)$ of IC model in [6], $\sigma_N(S, q)$ of IC-N model in [4], and $\sigma_t(S, T, \alpha)$ of t-IC model in [10]. Under t-IC-N model which is a mixture of IC-N model and t-IC model, it is trivial that $\sigma_{tN}(S, q, T, \alpha)$ also satisfies the three properties.

4. IPA ALGORITHM

In this section, we propose IPA algorithm which approximates influence spread for a series of IC models introduced in Section 2. Even though the NP-hardness is approximated by Greedy algorithm, the influence maximization problem still suffers from scalability challenge – evaluating $\sigma(S)$. $\sigma(S)$ is #P-hard which is intractable in a polynomial time [2]. The intuition behind the #P-hardness is that we cannot count all the path from seed nodes to non-seed nodes. To deal with the #P-hardness of evaluating $\sigma(S)$, existing works use Monte-Carlo simulations [1, 6], break the whole graph into several communities [12], or exploit local influence structures [2, 9].

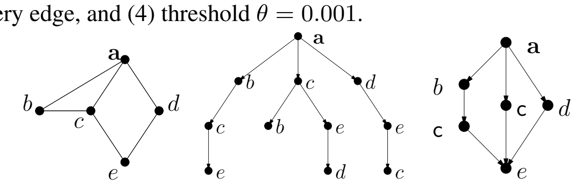
IPA [8] is an order of magnitude faster algorithm, which approximates $\sigma(S)$ without sacrificing influence spread, than the state of the art PMIA [2]. The main idea of IPA is that (1) an influence path from a seed node to a non-seed node is considered as an influence evaluation unit, (2) each influence path is mutually independent, and (3) a threshold θ controls the number of significant influence paths. IPA decompose approximation of $\sigma(S)$ into approximating (1) $\sigma(\{v\})$ of a single node and (2) $\Delta(S,v) = \sigma(S \cup \{v\}) - \sigma(S)$ of marginal influence spread increase. From now on, let $\hat{\sigma}(S)$ denote the approximation of $\sigma(S)$. [8] provides a detailed explanation of IPA.

For ease of understanding, we first figure out how IPA works under IC model – evaluating $\hat{\sigma_I}(\{v\})$ and $\hat{\Delta}_I(S,v)$. To evaluate $\hat{\sigma_I}(S)$, IPA first collects all influence paths which have influence propagation probability no less than θ. A sequence of nodes $\langle v_1, \cdots, v_m \rangle$ represents an influence path p. Then, the influence propagation probability of p is

$$ipp_I(p) = \begin{cases} 0 & , p = \langle\rangle \\ \prod_{i=1}^{m-1} w_{(v_i, v_{i+1})} & , p = \langle v_1, \cdots, v_m \rangle \end{cases} . \quad (2)$$

IPA collects influence paths by repeatedly expanding the outgoing edge of each existing path's ending node. A path expansion stops when the expanded path becomes a cycle or its $ipp_I(\cdot)$ is less than θ. The following example shows how IPA gathers influence paths staring from a node.

Example 1. [Influence path collection starting from $\mathbf{a}$]
Suppose that (1) a graph is Figure 1a, (2) influence diffusion model is IC, (3) propagation probability of 0.1 is uniformly assigned to every edge, and (4) threshold $\theta = 0.001$.



(a) Graph (b) Path expansion tree of $\mathbf{a}$ (c) Paths from $\mathbf{a}$ to e

Figure 1: Path collection starting from $\mathbf{a}$

All ten paths from the root $\mathbf{a}$ to the other nodes in Figure 1b are influence paths starting from $\mathbf{a}$. Also, Figure 1c shows that even though two influence paths shares a node c they are independent.

IPA evaluates $\hat{\sigma_I}(\{v\})$ of a single node using the collected influence paths staring from v. Let $P_{v \to u} = \{p | p = \langle v, \cdots, u \rangle\}$, $P_{S \to v} = \{p | p = \langle s, \cdots, v \rangle, s \in S\}$, $O_v = \{u | \langle \cdots, u \rangle \in P_{v \to V}\}$, and $O_S = \{u | \langle \cdots, u \rangle \in P_{S \to V}\}$. Then, the approximated influence spread of v is

$$\hat{\sigma_I}(\{v\}) = 1 + \sum_{u \in O_v} \hat{\sigma_I^u}(\{v\}). \quad (3)$$

$\hat{\sigma_I^u}(\{v\})$ is the approximated influence spread from v to u, and is the complement of the probability that no path in $P_{v \to u}$ activates u.

$$\hat{\sigma_I^u}(\{v\}) = 1 - \prod_{p \in P_{v \to u}} (1 - ipp_I(p)). \quad (4)$$

$\hat{\Delta}_I(S,v)$ evaluation is more complex than $\hat{\sigma_I}(\{v\})$. The difficulty of evaluating $\hat{\Delta}_I(S,v)$ is that the influence blocking illustrated in Figure 2 makes $\hat{\Delta}_I(S,v) \neq \hat{\sigma_I}(\{v\})$. Accordingly, we should filter out blocked(invalid) influence paths to evaluate $\hat{\Delta}_I(S,v)$.

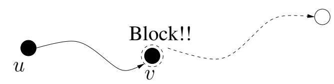


Figure 2: A situation that a seed node v blocks the influence of the other seed node u. (black circle : seed node, white circle : non-seed node)

Let $P_{S \to u}^{vaild}$ denote valid influence path set from seed nodes to $u \in V \backslash S$:

$$P_{S \to u}^{valid} = \{p | p \in P_{S \to u}, |p \cap S| = 1\}. \quad (5)$$

Then, the approximated marginal influence spread increase is

$$\hat{\Delta}_I(S,v) = 1 + \sum_{u \in O_v \cup \{v\}} \hat{\Delta}_I^u(S,v). \quad (6)$$

The range of the summation in Equation 6 shrinks from $O_{S \cup \{v\}}$ to $O_v \cup \{v\}$ by considering the common influence paths of $\{p | p = \langle u, \cdots, w \rangle, u \in S, w \notin O_v \cup \{v\}\}$. $\hat{\Delta}_I^u(S,v) = \hat{\sigma_I}(S \cup \{v\}) - \hat{\sigma_I}(S)$ is the marginal influence increase of v that affects u. $\hat{\sigma_I}(S)$ is influence spread from S to u as follows:

$$\hat{\sigma_I^u}(S) = 1 - \prod_{p \in P_{S \to u}^{valid}} (1 - ipp_I(p)). \quad (7)$$

Along with above description of IPA under IC model, IPA is easily adaptable to other extensions of IC model – IC-N, t-IC, and t-IC-N models. Because IC-N, t-IC, and t-IC-N models depend on independent one-to-one influence propagation, the framework of IPA need not be changed. Accordingly, (1) defining $ipp(p)$ and (2) replacing the term 1 of Equations 3 and 6 with q for IC-N and t-IC-N models are sufficient to approximate the influence approximation. Positive influence propagation probability under IC-N model is

$$ipp_N(p) = q^m \cdot \prod_{i=1}^{m-1} w_{(v_i, v_{i+1})} = q^m \cdot ipp_I(p) \quad (8)$$

, with quality factor q. Under t-IC model, a matrix $\mathbf{C}_{uv}$ embeds all possible propagation probability of an edge (u,v) until time limit T.

$$\mathbf{C}_{uv} = \begin{pmatrix} 0 & c_{uv}^{(1,0)} & \cdots & c_{uv}^{(T-1,0)} & c_{uv}^{(T,0)} \\ 0 & 0 & \cdots & c_{uv}^{(T-1,1)} & c_{uv}^{(T,1)} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & c_{uv}^{(T,T-1)} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad (9)$$

where $c_{uv}^{(t,i)} = w_{(u,v)} \cdot \delta(t, i-1) \prod_{j=0}^{t-i-2} (1 - w_{(u,v)} \cdot \delta(t,j))$ which means the probability that u activate v at ith trial. Then, the influence propagation probability of a path under t-IC model is

$$ipp_t(p) = (1\ 0 \cdots 0) \Big(\prod_{i=0}^{m-1} \mathbf{C}_{v_i v_{i+1}} \Big) (1\ 1 \cdots 1)^{\mathsf{T}}. \quad (10)$$

Same to the relationship between $ipp_I(\cdot)$ and $ipp_N(\cdot)$, positive influence propagation probability under t-IC-N model is

$$ipp_{tN}(p) = q^m \cdot ipp_t(p). \quad (11)$$

In addition to time efficient processing, IPA handles memory efficiently by throwing out insignificant influence paths, and is also parallelizable due to its independence between influence paths. The details are omitted due to the space limit and they are described in [8].

5. EXPERIMENT

Table 1: The basic statistics of the datasets

Dataset	Epinion	Stanford	DBLP	Patent	LiveJournal
# of Nodes	75.8K	281K	655K	3.77M	4.85M
# of Edges	509K	2.31M	3.98M	16.5M	69.0M

In this section, we briefly report processing time and influence spread of resultant seed nodes of IPA and its four competitors – PMIA,Greedy,SD and Random – on five datasets under IC model. Table 1 shows basic information on five dataset. More detailed experimental setup and results on memory efficiency and parallelization effect are available in [8].

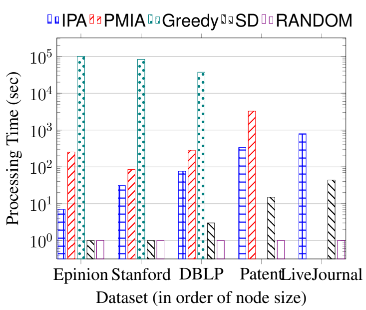


Figure 3: Processing Time

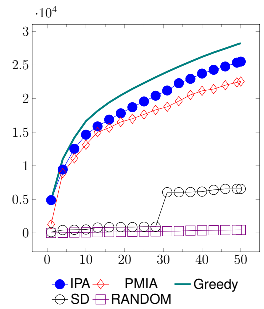


Figure 4: Influence spread on Stanford

Figure 3 illustrates the processing time of five influence maximization algorithms with log-scaled y-axis. We omit the processing time which is larger than 10^5 seconds. IPA shows an order of magnitude less processing time than PMIA. In addition, IPA has monotonically increasing processing time in proportion to the number of nodes, while the processing time of PMIA and Greedy fluctuates. SD is fast because it does not consider influence diffusion.

Figure 4 illustrates the influence spread of each algorithm's solution seed nodes on Stanford dataset. Obviously, Greedy is the best, but trades scalability for effectiveness. On the contrary, SD trades effectiveness for scalability. Among IPA and PMIA, IPA shows more influence spread. This trends are similar on other four datasets and all extended IC models.

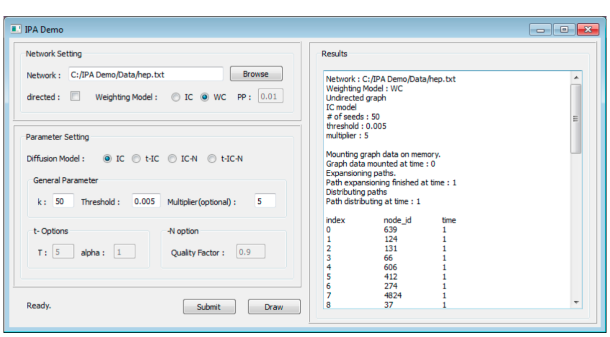


Figure 5: IPA desktop application

6. DEMONSTRATION

We provide a desktop application which finds seed nodes using IPA. The application is implemented using C++ with QT framework and Graphviz library. The application is available at `http://dm.postech.ac.kr/ipa_demo`

Figure 5 shows the user interface of the application. After setting a graph dataset, the number of seed nodes, a threshold, information diffusion model, and parameters of influence diffusion model, we have the resultant seed nodes, their ids, and their pop-up times in the result text area of the application.

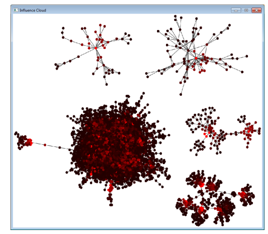


Figure 6: Influence Cloud

In addition, after getting the resultant seed nodes, the application visualizes a sub-graph of the target dataset which includes all valid influence paths. Figure 6 shows an influence sub-graph of 50 seed nodes of NetHEPT dataset. The color of a node represents the expectation value of each node being influenced. A node which have high expectation value has redder color.

References

[1] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 199–208, 2009. ACM.

[2] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1029–1038, 2010. ACM.

[3] W. Chen, Y. Yuan, and L. Zhang. Scalable influence maximization in social networks under the linear threshold model. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, pages 88–97, 2010. IEEE Computer Society.

[4] 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*, pages 379–390, 2011.

[5] L. Dagum and R. Menon. Openmp: An industry-standard api for shared-memory programming. *IEEE Comput. Sci. Eng.*, 5:46–55, January 1998.

[6] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146, 2003. ACM.

[7] D. Kempe, J. M. Kleinberg, and É. Tardos. Influential nodes in a diffusion model for social networks. In L. Caires, G. F. Italiano, L. Monteiro, C. Palamidessi, and M. Yung, editors, *ICALP*, volume 3580 of *Lecture Notes in Computer Science*, pages 1127–1138. Springer, 2005.

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

[9] M. Kimura and K. Saito. Tractable models for information diffusion in social networks. In J. Fürnkranz, T. Scheffer, and M. Spiliopoulou, editors, *Knowledge Discovery in Databases: PKDD 2006*, volume 4213 of *Lecture Notes in Computer Science*, pages 259–271. Springer Berlin / Heidelberg, 2006.

[10] W. Lee, J. Kim, and H. Yu. Influence maximization for time-considering independent cascade model. Technical Report 2012-02-tIC, Pohang University of Science and Technology (POSTECH), 2012. URL `http://dm.postech.ac.kr/techreport/TechReport-POSTECH-CSE-2012-02-tIC.pdf`.

[11] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 420–429, 2007. ACM.

[12] Y. Wang, G. Cong, G. Song, and K. Xie. Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 1039–1048, 2010. ACM.